Bioinformatische Analyse und funktionelle Charakterisierung von strukturellen Genvarianten in ADME-Genen in humaner Leber

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

Fakultät Naturwissenschaften

Universität Hohenheim

Prof. Dr. Ulrich M. Zanger

Dr. Margarete Fischer-Bosch-Institut für klinische Pharmakologie

Prof. Dr. Lutz Graeve

Institut für Biologische Chemie und Ernährungswissenschaft,

Universität Hohenheim

vorgelegt von

Roman Tremmel

aus Ostfildern-Ruit

Hohenheim, 2016

Dekan: Prof. Dr. Heinz Breer

1. berichtende Person: Prof. Dr. Ulrich M. Zanger

2. berichtende Person: Prof. Dr. Lutz Graeve

Eingereicht am: 03.02.2016

Mündliche Prüfung am: 14.03.2016

Die vorliegende Arbeit wurde am 11.03.2016 von der Fakultät Naturwissenschaften der Universität Hohenheim als
„Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften“
angenommen.

Für meine Familie

Kurzfassung

Die Pharmakogenetik beschreibt die Untersuchung von interindividuellen genetischen Unterschieden, welche die Wirkung von Medikamenten und die Reaktion auf xenobiotische Substanzen beeinflussen. Ein Großteil dieser Variabilität beruht auf Unterschieden im hepatischen Arzneimittelmetabolismus. Die daran beteiligten Enzyme, Transporter und Rezeptoren sind in die Prozesse der Absorption, der Distribution, des Metabolismus und der Exkretion involviert und werden als ADME-Gengruppe zusammengefasst. Neben den genetischen Faktoren können auch Einflüsse aus der Umwelt und endogene Faktoren, wie das Geschlecht, das Alter, Entzündungsreaktionen und andere Faktoren die Expression und die Aktivität der ADME-Gene ändern. Sind die Einflüsse so groß, dass das therapeutische Fenster eines Medikaments verlassen wird, kann das Ansprechen auf das Medikament ausbleiben oder unerwünschte Arzneimittelwirkungen und gegebenenfalls Toxizität können auftreten.

Das humane Genom eines bestimmten Individuums unterscheidet sich neuesten Daten zufolge im Vergleich zum Referenzgenom an 4,1 bis 5 Millionen Positionen. Davon sind mehr als 99,9% Einzelbasenpolymorphismen (*single nucleotide polymorphism*; SNP) oder kleine Insertionen und Deletionen. Daneben kommen bis zu 2.500 strukturelle Varianten pro Person vor, die insgesamt ca. 20 Millionen Basen einschließen und somit auf DNA-Basis mehr zur Variabilität des Genoms beitragen als die SNPs (1000 Genome Project, 2015). In der Gruppe der strukturellen Varianten finden sich auch Kopienzahlvariationen (*copy number variations*; CNV), per Definition mindestens 1kb lange duplizierte oder deletierte DNA Segmente. Beobachtungen zur Funktionalität der CNVs in der Fruchtfliege, der Maus und im Menschen fanden neben CNVs, die einen Einfluss auf den Phänotyp zeigten (dosis-sensitiv), auch CNVs, deren Einfluss umgekehrt (dosis-umgekehrt) oder komplett abwesend war (dosis-insensitiv). CNVs, die den Arzneimittelmetabolismus beeinflussen, wurden zum Beispiel für das Phase I Gen *CYP2D6* gezeigt. Eine verminderte oder erhöhte Kopienzahl des Gens wirkt sich direkt sowohl auf die mRNA und Proteinexpression, als auch auf die Enzymaktivität aus. So ist der Metabolismus der CYP2D6 Substrate Codein (Opioid) oder Tamoxifen (selektiver Östrogenrezeptormodulator) in Trägern dieser Varianten im Vergleich zu Trägern mit zwei Kopien signifikant verändert. Eine im Vorfeld durchgeführte Genotypisierung kann deswegen in Zukunft helfen, die Medikamentendosis Genotyp-abhängig anzupassen.

Diese Arbeit beschäftigt sich mit der Fragestellung, ob weitere ADME-Gene CNVs aufweisen und ob diese genetischen Varianten einen funktionellen Einfluss auf den Phänotyp und Arzneimittelmetabolismus haben.

Dazu wurde systematisch das Vorkommen von CNVs der wichtigsten ADME-Gene (n=340) in drei unabhängigen Datensätzen untersucht. In einer öffentlichen Datenbank für genomische strukturelle Varianten (DGV; dgv.tcag.ca) wurde die Position der ADME-Gene mit den dort beschriebenen CNV-Positionen aus HapMap-Proben abgeglichen. Zusätzlich wurden in gesunden (n=269) und malignen (n=351) humanen Leberproben des TCGA-Projektes (http://cancergenome.nih.gov/) prozessierte SNP-Mikrochipdaten ausgewertet. In einer am IKP Stuttgart etablierten Biobank (IKP148), die 150 Leberproben mit europäischer Herkunft und vollständiger klinischer und demographischer Dokumentation umfasst, wurde mit Hilfe eines ADME-panelbasierten *next generation* Exonsequenzierungsprojektes (NGS) die Kopienzahl der 340 ADME-Gene detektiert. Dafür wurde eine Methode entwickelt und ein bioinformatischer Arbeitsablauf angewendet, der die Abdeckung der einzelnen Sequenzierungen auswertet und daraus die relative Kopienzahl für jedes Gen und Exon bestimmt. Die Ergebnisse wurden mittels *real time* PCR und spezifischen TaqMan CNV-Assays validiert. Für die funktionelle Assoziationsanalyse wurde in 50 gesunden TCGA-Leberproben auf Genexpressionswerte einer RNA-Sequenzierung zurückgegriffen. In lymphoblastoiden Zelllinien der HapMap-Proben (LCL) und den IKP148-Leberproben wurden normalisierte Intensitäten eines Expressionsmikrochips zur funktionellen Analyse verwendet. Zusätzlich lagen in den Leberproben (IKP148) bereits Protein und Enzymaktivitätsdaten für einzelne ADME-Gene vor. Weitere Proteinexpressionsdaten wurden mit WesternBlot Analysen bestimmt.

Mit diesem systematischen Ansatz konnten alle bekannten, pharmakologisch wichtigen CNVs in Phase I und II Genen, wie *CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT1*, *SULT1A1* und *UGT2B17* in allen untersuchten Datensätzen bestätigt werden. Weitere CNVs, die teilweise bereits bekannt waren oder noch nicht als funktionell beschrieben wurden, fanden sich ebenfalls in der Phase I und II Gruppe, wie *CES1, CYP2E1, CYP21A2, UGT2B15* und *UGT2B28.* Seltene CNVs (<1%) betrafen überwiegend Transportergene, wie zum Beispiel *ABCA2*, *SLC2A4* und *SLC47A1*. Zudem wurden in den Leberproben (IKP148) mit der NGS-Methode *CYP2A6* und *CYP2D6* CNVs feinkartiert. Mit einer Auflösung auf Exonebene wurden Hybridallele der beiden Gene mit dem jeweiligen Pseudogen festgestellt und bestätigt. Auch die funktionelle Analyse bestätigte die positive Assoziation von CNVs der Gene *CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT1*, *SULT1A1* und *UGT2B17* mit der mRNA Expression in allen drei Kohorten. In den IKP148-Leberproben wurde mit der Kombination aller genetischen Informationen aus dem

ADME-panelbasierten NGS-Datensatz gezeigt, dass die genetischen Faktoren 11% bzw. 53% der Variabilität der CYP2A6 und CYP2D6 Enzymaktivität erklären.

Im Gegensatz dazu war die mRNA Expression der Gene *CES1* oder *CYP2E1* in gesundem Lebergewebe (IKP148 und TCGA) und in den LCLs nicht von der Kopienzahl abhängig. Eine detailliertere Analyse des *CYP2E1* Gens und der detektierten CNVs mit der Proteinexpression und Enzymaktivität bestätigte die Unabhängigkeit von der Gendosis in den IKP148-Leberproben. Diese Dosiskompensation kann prinzipiell durch verschiedene Mechanismen erklärt werden. Neben gewebe- und tumorspezifischen Faktoren könnten gelinkte genetische Varianten, eine starke posttranskriptionelle und posttranslationale Regulation wie z.B. mit Hilfe von miRNAs, ein unvollständiger Einschluss der regulatorischen und kodierenden Sequenzen in der strukturellen Variante, Hybridgene, eine monoallelische Expression, negative Rückkopplung oder epigenetische Faktoren für das Ausbleiben des CNV-Effektes verantwortlich sein. In dieser Arbeit wurde mittels einer Haplotypanalyse in der *CYP2E1* Region SNPs identifiziert, die mit der Genduplikation gelinkt waren und tendenziell die Expression in Individuen mit europäischer Herkunft negativ beeinflussten. Mit *in silico* Methoden wurde eine Beziehung zwischen einem der zur Duplikation gelinkten SNPs in der 3'UTR mit zusätzlichen miRNA Bindestellen vorhergesagt. Die daran bindenden miRNAs sind möglicherweise für die Regulation zusätzlicher Kopien verantwortlich.

Der Einfluss von CNVs auf die mRNA Expression anderer Gene, wie *CYP21A2*, *UGT2B25* und *UGT2B28*, war inkonsistent. Obwohl *CYP21A2* Deletionen mit einem verminderten Expressionsphänotyp assoziiert waren, zeigten Duplikationen keinen Einfluss (IKP148). Eine Assoziation von *UGT2B28* CNVs wurde nur in LCLs nicht aber in Lebergewebe (IKP148 und TCGA) gefunden. Insgesamt wurde in den IKP148-Proben in 7 von 17 Genen, in den TCGA-Leberproben in 2 von 12 und in den LCLs in 3 von 14 Genen ein Einfluss der CNVs auf die Expression gefunden. Im TCGA-Krebsgewebe waren nahezu alle 340 ADME-Gene von CNVs betroffen und mehr als 30% der CNVs zeigten einen signifikanten Effekt auf die mRNA Expression.

Im Rahmen von Kooperationsprojekten wurden weitere Polymorphismen und Phänotypen von CYP2E1 und SULT1A1 analysiert.

CYP2E1: Dieser Teil der Arbeit befasst sich mit Faktoren, die das Risiko einer Erkrankung an Schilddrüsenkrebs (*differentiated thyroid carcinoma*; DTC) verändern. Es ist bekannt, dass die Progression von DTC durch das Zusammenspiel von genetischen Faktoren, Umwelteinflüssen, wie ionisierende Strahlung, vorhergegangene Erkrankungen der Schilddrüse und unnatürliche Hormonspiegel, beeinflusst wird. Acrylamid, das mit der Nahrung aufgenommen wird, gilt als potentieller neuer Faktor in der DTC Ent-

stehung. Da Acrylamid von CYP2E1 zum karzinogenen Glycidamid metabolisiert wird und diese Reaktion vermutlich abhängig vom *CYP2E1* Genotyp ist, wurde untersucht, ob Varianten von *CYP2E1* das DTC Risiko beeinflussen können. Dafür wurde in einer Fall-Kontroll-Studie eines Kooperationspartners (Prof. Dr. Stefano Landi, Universität von Pisa, Pisa, Italien) ein Tag-SNP Verfahren gewählt und der SNP rs2480258, der Varianten im Intron acht und der 3'UTR von CYP2E1 abdeckt, als Risiko SNP identifiziert. Anschließend wurde untersucht, ob der Tag-SNP die Expression und Aktivität von CYP2E1 in humanem Gewebe beeinflusst. Dazu wurden in den Leberproben (IKP148) die Genotypen des tag-SNP durch ein Imputierungsverfahren bestimmt und mit den CYP2E1 Phänotypen korreliert. Es zeigte sich, dass das A Allel des tag-SNPs, welches mit einem höheren DTC-Risiko assoziiert war, die mRNA, die Proteinexpression und die Enzymaktivität reduziert. Eine *in silico* Analyse zum möglichen molekularen Mechanismus prognostizierte, dass eine miRNA (miR570) die CYP2E1 Transkripte in Trägern des A Allels reduziert. Dieses Ergebnis deutet darauf hin, dass die interindividuelle CYP2E1 Aktivität sowie Acrylamid (ähnlich dem Glycidamid) das Risiko für die Entstehung von DTC beeinflussen.

SULT1A1: Methyleugenol, ein Sekundärmetabolit aus Kräutern wie Lorbeer und Basilikum, wird im humanen Organismus über eine Sulfatierung zu einem reaktiven Metaboliten aktiviert, welcher DNA kovalent binden kann. Die entstehenden DNA-Addukte sind mutagen und können bei ineffektiver DNA-Reparatur ein Auslöser von Karzinogenese sein. In Mäusen wurde von einem Kooperationspartner (Prof. Dr. Hans-Rudolf Glatt, Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke (DIfE), Nuthetal, Deutschland) bereits gezeigt, dass der Metabolismus in der Leber abläuft und hauptsächlich das Phase II Enzym SULT1A1 daran beteiligt ist. Um diese Ergebnisse in humanen Proben nachzuvollziehen, wurden vom Kooperationspartner in 121 der 150 Leberproben (IKP148) die Methyleugenol DNA-Adduktspiegel gemessen. In dieser Arbeit wurde zusätzlich die SULT1A1 Proteinmenge in den IKP148-Leberproben bestimmt. Die SULT1A1 mRNA und Proteinexpression korrelierten jeweils signifikant mit den DNA-Adduktmessungen, d.h. je höher die SULT1A1 Expression, desto höher waren die Adduktspiegel. Dies bestätigte in humanen Leberproben die *in vivo* Rolle von SULT1A1 im Methyleugenolmetabolismus. Wie oben bereits angedeutet, kamen in den Leberproben Träger von ein, zwei, drei, vier und fünf *SULT1A1* Genkopien vor. Dabei waren Deletionen seltener (4%) als Duplikationen (36%) zu beobachten. Die Kopienzahl korrelierte signifikant mit der SULT1A1 mRNA und Proteinexpression. Diese Ergebnisse sind in Einklang zu früheren Studien, die einen Effekt der Kopienzahl auf die Enzymaktivität aufzeigten. Die Methyleugenol DNA-Adduktspiegel waren ebenfalls zur *SULT1A1* Kopienzahl assoziiert. Träger von mehr als drei Genkopien hatten ein 2,8-

fach höheres DNA-Adduktlevel als Träger mit nur einer *SULT1A1* Genkopie. Träger mehrerer *SULT1A1* Kopien können daher unter Umständen leichter, schneller und öfter ein kritisches DNA-Adduktlevel erreichen, das zu einem höheren Krebsrisiko führen kann. Weiterführende Fall-Kontroll-Studien sollten daher, neben der Menge an aufgenommenem Methyleugenol, die *SULT1A1* Kopienzahl berücksichtigen.

Schlagwörter: Pharmakogenetik, Pharmakogenomik, Kopienzahlvariation, CNV, humane Leber, Arzneimittelmetabolismus, dosis-sensitiv, dosis-insensitiv, Expression, Phänotyp, Assoziation, NGS, Exonsequenzierung

Abstract

Pharmacogenetics is the study about inter-individual genetic variation that influences the response to drugs and other xenobiotics. A major part of this variation is due to hepatic drug metabolism with enzymes, transporters and receptors involved in the absorption, distribution, metabolism and the excretion of drugs, xenobiotics and endogenous substances and collectively defined as ADME-genes. Genetic factors along with environmental and endogenous factors, including gender, age, inflammation processes and others are known to influence the expression and activity of ADME-genes. These influences can affect drug response, side effects or toxicity.

According to newly published data, the human genome of any subject differs from a reference genome at 4.1 to 5.0 million positions. More than 99.9% of these differences are single nucleotide polymorphisms (SNP) or short insertions or deletions. Furthermore, a person carries up to 2,500 structural variants, including copy number variations (CNV) affecting ~20 million bases (1000 Genomes Project Consortium et al., 2015). Thus structural variants affect more bases than SNPs. Per definition the CNVs are duplicated or deleted DNA segments greater than 1kb and it was shown that they cover at least 12-30% of the human genome. Genome-wide studies investigating the functionality of CNVs in the fruit fly, the mouse and in humans showed that there are genes whose expression is clearly affected by CNVs (dosage-sensitve), but also genes showing lower expression with increased copy number (dosage reversed) or genes without any expression alterations despite different copy number (dosage-insensitive).

A prominent example of CNVs influencing drug metabolism is the phase I gene *CYP2D6*. Carriers of reduced or amplified gene copies show significantly altered expression and enzyme activity levels and also a different drug metabolism of substrates like codeine (opioid) or tamoxifen (selective estrogen receptor antagonist) in comparison to carriers with normal copy status of two. Genotyping of *CYP2D6* gene copy number may thus help to adjust drug dosage in a genotype dependent manner. In this work I investigated if further ADME-genes are affected by CNVs and if these variants have a functional impact on the expression phenotype and drug metabolism.

The distribution of CNVs in the most important ADME-genes (n=340) was investigated in three independent cohorts using CNV data in a public accessible database of genomic variants (DGV; dgv.tcag.ca), processed SNP microarray data of paired samples of healthy (n=269) and tumor (n=351) liver tissue of the TCGA project (http://cancergenome.nih.gov/) and ADME-panel based exon next generation sequencing (NGS) applied on 150 well documented human liver samples of an in-house cohort (IKP148). For the NGS data analysis a method was developed and optimized to esti-

mate the relative copy number of the ADME genes or every single exon via the read depth. The results were validated using qPCR with specific TaqMan assays. RNA-sequencing data of 50 healthy TCGA liver samples, and normalised expression data from microarray experiments applied to lymphoblastoid cell lines (LCL) from the HapMap samples and the 150 human liver samples (IKP148) were used to analyse the association between CNVs and the mRNA expression. Furthermore, in the IKP148 liver samples protein and enzymatic activity levels were available or measured using WesternBlot and mass spectrometry for selected ADME-genes.

All pharmacologically important CNVs of phase I and phase II genes, including *CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT1*, *SULT1A1* and *UGT2B17* could be confirmed in all datasets. CNVs which were known, but so far not functionally assessed were found in the phase I and II genes *CES1, CYP2E1, CYP21A2, UGT2B15* and *UGT2B28*. In this work rare CNVs (<1%) were mainly found for transporters like *ABCA2*, *SLC2A4* and *SLC47A1*. The analysis of the read depth in the IKP148 samples data revealed hybrid genes for *CYP2A6* and *CYP2D6* with their pseudogenes and allowed a fine mapping of the different alleles. The functional analysis further confirmed the positive association between CNVs and the mRNA expression of *CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT1*, *SULT1A1* and *UGT2B17* in all three cohorts. The combination of all data from the NGS project in the IKP148 liver subjects, including SNP and CNV genotypes showed that 11% and 53% of the variability of CYP2A6 and CYP2D6 enzyme activity were explained by the genetic factors.

In contrast the mRNA expression of the genes *CES1* and *CYP2E1* was not dependent of the CNV pattern in healthy liver tissue (IKP148 and TCGA) and lymphoblastoid cell lines. A detailed analysis of the protein and enzyme activity levels (chlorzoxazone-6-hydroxylation) of CYP2E1 confirmed the dosage-insensitivity in the IKP148 liver subjects. The dosage compensation can be principally explained by different mechanisms and could be tissue or tumor specific. Furthermore, CNV-linked genetic variants, altered miRNA regulation, incomplete inclusion of regulatory elements or coding sequences, hybridgenes, monoallelic expression, feedback loops or epigenetics could be factors which mask the CNV effect. In this work a haplotype analysis of the *CYP2E1* region identified SNPs which were linked to the duplication and a reduced expression phenotype in persons with European ancestry. Using *in silico* prediction tools we found a relation of one of the linked SNPs in the 3'UTR with additional predicted miRNA binding sites potentially regulating additional *CYP2E1* gene copies.

The CNV influence on the mRNA expression of the genes *CYP21A2*, *UGT2B25* and *UGT2B28* was inconsistent. Although *CYP21A2* deletions were associated with a decreased expression, gene duplications showed normal expression levels compared to

samples with two copies. A significant influence of *UGT2B28* CNVs was found in LCLs but not in human liver samples (IKP148 and TCGA). In total 7 of 17, 2 of 12 and 3 of 14 ADME genes showed a significant association between expression and CNV type in the IKP148, TCGA and LCLs of the HapMap samples, respectively. In the TCGA cancer tissue nearly all ADME-genes carry CNVs and in 30% of the genes a significant correlation was observed.

With cooperation partners further polymorphisms and phenotypes of SULT1A1 and CYP2E1 were analyzed.

CYP2E1: In this part of the thesis factors influencing the risk of developing differentiated thyroid carcinoma (DTC) were investigated. Known risk factors for the progression of DTC are genetic and environmental factors, including ionizing radiations, previous thyroid diseases, and hormone factors. It has been speculated that dietary acrylamide intake correlates with the DTC formation. The acrylamide molecule is metabolized by CYP2E1 to the reactive carcinogenic glycidamide. The enzymatic reaction is probably dependent on the CYP2E1 genotype. Together with a cooperation partner (Prof. Dr. Landi, University of Pisa, Pisa, Italy) we investigated, whether *CYP2E1* variants influence the DTC risk. Prof. Landi and colleagues used a case-control-cohort and a haplotype approach and observed a significant association between a tag-SNP rs2480258 (A allele), which covers variants in intron eight and the 3'UTR, and an increased DTC risk. In the human liver samples (IKP148) the rs2480258 genotypes were assessed using an imputation analysis and it was shown that particularly the A allele of the SNP reduce significantly the mRNA and protein expression and the enzyme activity. An *in silico* prediction for the molecular mechanism suggested that miR570 specifically down regulates the transcripts in carriers of the A allele. These results indicated that the inter-individual CYP2E1 activity as well as acrylamide (similar to glycidamide) influences the risk for DTC.

SULT1A1: Methyleugenol, a secondary metabolite present in herbs such as basil or laurel is metabolized in humans by sulfation to a reactive product which can covalently bind to DNA. The resulting DNA adducts are mutagenic and can promote carcinogenesis. Our cooperation partner Prof. Dr. Hans-Rudolf Glatt (German Institute for Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany) had shown, that the methyleugenol metabolism takes place in the liver of mice and is mainly catalyzed by the phase II enzyme SULT1A1. To investigate these facts in humans, the methyleugenol DNA-adduct levels were measured by the cooperation partner in liver tissues (n=121; IKP148) using mass spectrometry. In this work the SULT1A1 protein levels were determined using western blot analysis and the relation between the DNA adducts as well as the SULT1A1 expression and the *SULT1A1* CNVs was assessed. The SULT1A1

mRNA and protein expression were significantly correlated to the DNA adducts, e.g. higher SULT1A1 expression resulted in higher adduct levels. This emphasized the role of SULT1A1 in the *in vivo* metabolism in human liver samples. As mentioned above, there were individuals (IKP148) carrying one, two, three, four and five copies of *SULT1A1*. Deletions were found less frequent (4%) than duplications (36%). The CNVs were significantly associated with the SULT1A1 mRNA and protein expression. This result was consistent to previous studies investigating the association between *SULT1A1* CNVs and enzyme activity. The methyleugenol DNA adduct levels were also significantly associated to the SULT1A1 copy number. Carriers of at least three gene copies exhibited a 2.8-fold higher DNA adduct level compared to donors carrying only one *SULT1A1* gene copy. As a consequence this could mean that individuals with multiple *SULT1A1* copies reach faster, more often and more easily critical and ultimate adduct levels which increase the risk for developing cancer. Future studies should clarify whether methyleugenol intake as well as the individual *SULT1A1* CNV make-up influences the risk of cancer.

Keywords: pharmacogenetics, pharmacogenomics, copy number variation, CNV, human liver, drug metabolism, gene dosage, expression, phenotype, association, NGS

Inhaltsverzeichnis

Abkürzungsverzeichnis

ABC	*ATP binding cassette*
ADME	Absorption, Distribution, Metabolismus und Exkretion
bp	Basenpaar
CNV	*Copy number variation*; Kopienzahlvariation
CNVR	CNV Region
CRP	C-reaktives Protein
CYP	Cytochrom P450
DDI	*Drug-drug interactions* ; Medikamentinteraktionen
DGV	*Database of genomic variants;* Datenbank für genomische strukturelle Varianten
DNA	Desoxyribonukleinsäure
FC	*Fold change*; x-facher Unterschied
GO	Genontologie
GST	Glutathion S-Transferase
HCC	Hepatozelluläres Karzinom; Leberkarzinom
HWE	Hardy-Weinberg-Equilibrium
Indel	*Small insertion-deletion*; Kleine Insertion-Deletion
LCL	Lymphoblastoide Zelllinie
LC-MS/MS	Flüssig-Chromatographie-Massenspektometrie/Massenspektometrie
LD	*Linkage disequilibrium*; Kopplungsungleichgewicht
NGS	*Next generation sequencing*; Sequenzierung der nächsten Generation
PCA	*Principal component analysis*; Hauptkomponentenanalyse
qPCR	Quantitative Polymerase-Kettenreaktion
RNA	Ribonukleinsäure
OR	*Odds ratio*; Quotenverhältnis
ROS	*Reactive oxygen species*; Reaktive Sauerstoffspezies
SD	Segmentale Duplikation
Seq	Sequenzierung hier: NGS

sm	Mittelwert der Intensitäten im CNV-Segment
SNP	*Single nucleotide polymorphism;* Einzelbasenmutation
SULT	Sulfotransferase
TCGA	*The cancer genome atlas;* Krebsgenom Atlas
UGT	UDP-Glucuronosyltransferase
UTR	*Untranslated region;* Untranslatierter Bereich
VAF	*Variant allele frequency;* Variationsfrequenz

1 Einleitung

1.1 Grundlagen des Fremdstoffmetabolismus

Exogene Substanzen (Xenobiotika), die mit der Nahrung aufgenommen werden, wie sekundäre Pflanzenstoffe oder Medikamente, haben meistens lipophile Eigenschaften und können zwar leicht absorbiert, allerdings nur schwer renal oder hepatisch ausgeschieden werden und reichern sich dadurch in Zellmembranen und Fettgewebe an. Die entscheidende Aufgabe des Fremdstoffmetabolismus ist, die Polarität der Substanzen zu erhöhen und dadurch die Elimination über die Nieren oder über die Galle zu erleichtern. Eine Vielzahl von Substanzen verliert durch die Biotransformation ihre pharmakologische Wirkung und ist somit deaktiviert. Allerdings findet man auch Medikamente und andere Xenobiotika, deren Metaboliten erst aktiv oder sogar toxisch sind (Zanger, 2012). Die Wirkung eines Medikaments ist also stark vom Fremdstoffmetabolismus abhängig.

In der Leber findet der wesentliche Teil des Arzneimittelmetabolismus statt. In den Leberzellen, den Hepatozyten, sind für den Metabolismus eine Vielzahl von spezialisierten Enzymen und Transportern exprimiert. Bisher werden ungefähr 340 Gene als die wichtigsten Komponenten im Fremdstoffmetabolismus angenommen (www.pharmaadme.org; Meyer, 1996). Diese Gene sind bei Prozessen der Absorption, Distribution, Metabolisierung und der Elimination beteiligt und werden als ADME-Gene zusammengefasst. Der Metabolismus lässt sich in mindestens drei Phasen aufteilen. In der Phase I werden lipophile Substanzen modifiziert, indem sie hydroxyliert oder oxidiert werden. In der Phase II werden polare Gruppen konjugiert und anschließend wird das Produkt in Phase III ausgeschieden (Anzenbacher and Anzenbacherová, 2012). Hinzu kommt die Regulation des Systems, zu der Modifizierer der ADME-Genexpression oder der Biochemie der ADME Enzyme zählen (Zanger and Schwab, 2013).

Die Cytochrom P450 Enzyme (CYP) sind die wichtigsten Enzyme des **Phase I** Metabolismus. Im humanen Genom wurden 57 Gene und 58 Pseudogene der CYP-Familie zugeordnet und in 18 Familien gruppiert (Nelson et al., 2004). Viele CYPs übernehmen eine wichtige endogene Rolle in der Biosynthese von Steroidhormonen und Gallensalzen und so sind nur 10-15 CYPs der Familien CYP1, 2, und 3 für den Großteil (80%) des Fremdstoffmetabolismus verantwortlich (Zanger et al., 2008). Die am höchsten exprimierten Enzyme sind CYP3A4, CYP2C9, CYP1A2 und CYP2E1 (Achour et al.,

2014). Die Enzyme katalysieren eine Monooxidation, welche meistens zu einer Hydroxylierung des Substrates führt. Endet der Sauerstoff allerdings nicht im Substrat, spricht man von einer entkoppelten Reaktion, die zu reaktiven Sauerstoffspezies (*reactive oxygen species*, ROS) führt. Diese entkoppelte Reaktion ist in der CYP-Familie unterschiedlich stark ausgeprägt und eine Charakteristik des CYP2E1 Enzyms, was dessen Rolle in toxischen Prozessen erklärt (Caro and Cederbaum, 2004). Weitere Enzyme der Phase I sind Alkohol Dehydrogenasen oder Esterasen, zu denen die Familie der Carboxylesterasen (CES1-3) gehört.

Enzyme der **Phase II** erhöhen durch eine Konjugation mit diversen geladenen Substanzen weiter die Polarität und Wasserlöslichkeit des exogenen Substrats. Am häufigsten finden Reaktionen wie eine Sulfatierung, Glucuronidierung oder Glutathion-Konjugation statt. Die wichtigsten verantwortlichen Enzyme des Phase II Arzneimittelmetabolismus sind Glutathion S-Transferasen (GST), Sulfotransferasen (SULT) oder UDP-Glucuronosyltransferasen (UGT). Daneben sind noch verschiedene Methyltransferasen (TPMT) oder N-Acetyltransferasen (NAT) von Relevanz (Jancova et al., 2010).

Humane GST-Enzyme sind in die vier Klassen Alpha (GSTA1-A4), Mu (GSTM1-M5), Pi (GSTP1), Kappa (GSTK1) and Theta (GSTT1, GSTT2) eingeteilt und im Zytosol, Mitochondrium oder endoplasmatischem Retikulum exprimiert (Jancova et al., 2010). Als Substrat kommen alle Stoffe in Frage, die mit der Thiolgruppe des Glutathion reagieren können. Darunter fallen Xenobiotika, wie industrielle Zwischenprodukte, Pestizide, Chemotherapeutika und endogene Stoffe, wie Prostaglandine und ROS. Die biologische Funktion ist vielfältig und erstreckt sich über Detoxifizierungsreaktionen, Hormonsynthese und Bioaktivierung und Inhibierung von Xenobiotika (Hayes et al., 2005).

Bisher sind in Säugetieren 117 Gene der UGT-Superfamilie beschrieben worden (Familien UGT1-3 und UGT8). Deren grundlegende Funktion ist die Glucuronidierung von exogenen (Medikamente, chemische Karzinogene und Umweltschadstoffe) und von endogenen Substanzen (Bilirubin, Steroidhormone, Schilddrüsenhormone und Gallensalze) (Mackenzie et al., 2005; Riedmaier et al., 2010). Ungefähr 40-70% der in der Klinik eingesetzten Medikamente werden von einem der UGT-Enzyme glucuronidiert (Wells et al., 2004). Im Detail wird die Bindung von einem Glucuronsäuremolekül an ein freies nukleophiles O-, N-, S-, oder C-Atom des Substrates katalysiert. Das entstehende Produkt kann dann leicht über die Galle oder den Urin ausgeschieden werden (Jancova et al., 2010).

SULTs kommen im Zytosol oder membrangebunden vor. Insgesamt sind 76 Isoformen bekannt. Die im Zytosol exprimierten Enzyme der Familien SULT1-5 sulfonieren eine Bandbreite von endogenen Substanzen (Hormone, Neurotransmitter), Medikamenten

oder Xenobiotika (Strott, 2002). Neben der Biotransformation wird eine Vielzahl von Pro-Karzinogenen durch die Sulfonierung aktiviert. Die reaktiven Zwischenprodukte können mutagen sein und die DNA kovalent binden. Durch die enzymatische Reaktion wird eine Sulfonylgruppe des universalen Sulfatspenders PAPS (3′-Phosphoadenosin-5′-phosphosulfat) an geeignete Akzeptorgruppen des Substrates übertragen (Anzenbacher and Anzenbacherová, 2012). SULTs sind wahrscheinlich die wichtigsten Detoxifizierungsenzyme im humanen Fetus, da bisher keine fetale UGT Transkription gemessen wurde (Jancova et al., 2010). In adultem Lebergewebe ist SULT1A1 die am stärksten exprimierte Isoform der SULT Familie (Riches et al., 2009).

Medikamente, deren Metabolite und andere Xenobiotika werden aktiv in die Zelle oder aus der Zelle heraus transportiert. Zu den wichtigsten **Transportern** im Arzneimittelmetabolismus gehören Aufnahmetransporter der *solute carrier* Familie (SLC) sowie Efflux Transporter aus der *ATP-binding-casette* Familie (ABC). Ihre Expression in Epithelien der Blut-Hirn-Schranke, des Dünndarms, der Niere und Leber unterstreicht ihre wichtige Rolle für die Distribution und die Pharmakokinetik von einer Vielzahl an Medikamenten (International Transporter Consortium et al., 2010). Die 48 Mitglieder umfassende Familie der ABC Transporter ist im Transport von endogenen und exogenen Substanzen involviert. Dazu zählen Medikamente, Hormone, Lipide und andere Xenobiotika (Schinkel and Jonker, 2003). Zu den wichtigsten Transportern im ADME-System gehören *ABCB1* (MDR1; p-Glykoprotein), *ABCC1* und 2 (MRP1 und 2) und *ABCG2* (BCRP).

1.2 Einflüsse und Variabilität im Arzneimittelmetabolismus

Faktoren, welche die inter- und intraindividuelle Variabilität der Expression und Aktivität der ADME-Gene und somit auch zu einer Veränderung der Wirkung und Nebenwirkung von Medikamenten führen können, sind extrinsischer oder intrinsischer Natur. Zu den Einflüssen von außen zählen Interaktionen von Medikamenten (*drug-drug interactions*; DDI), der Konsum von Alkohol und Nikotin und Komponenten in der Nahrung. Intrinsische Faktoren umfassen Geschlecht und Alter, Übergewicht, Abstammung, Krankheiten und Organfehlfunktionen sowie genetische Faktoren (Huang and Temple, 2008).

Das Alter kann eine wichtige Rolle im Arzneimittelmetabolismus einnehmen. Der größte Unterschied findet sich zwischen fetalem und adultem Gewebe. Für CYPs (CYP2E1, CYP2C9 und CYP2C19) aber auch SULTs (SULT1A1) wurde gezeigt, dass die Expression erst neonatal ansteigt oder, wie für SULT1A3, abgeschaltet wird (Jancova et al., 2010; Koukouritaki et al., 2004; Vieira et al., 1996). In älteren Personen beeinflussen vor allem eine in dieser Gruppe häufig verabreichte multimedikamentöse Therapie

und physiologische Faktoren, wie ein unnormaler Blutdruck oder eine verminderte renale Funktion das ADME-System (Cotreau et al., 2005; Kinirons and O'Mahony, 2004). Auch das Geschlecht hat einen Einfluss auf den Metabolismus. Dabei wirken sich die zwischen den Geschlechtern unterschiedlichen Parameter, wie das Körpergewicht, das Plasmavolumen oder die Fettspeicher auf die Pharmakokinetik aus (Beierle et al., 1999; Gandhi et al., 2004). Ein direkter Expressionsunterschied zwischen den beiden Geschlechtern wurde für CYPs, UGTs und GST beschrieben. So sind CYP1A2, CYP3A4 und CYP7A1 (höher in Frauen) und CYP3A5, CYP27B1 und UGT2B15 (höher in Männern) unterschiedlich exprimiert (Gallagher et al., 2010; Hoensch et al., 2006; Wolbold et al., 2003). Die erhöhte mRNA Expression resultiert in einer erhöhten Enzymaktivität. Dadurch schreitet die Verstoffwechselung von CYP3A4 Medikamenten in Frauen schneller voran (Cotreau et al., 2005; Wolbold et al., 2003). In Männern wurde eine höhere CYP2E1 Aktivität festgestellt, die allerdings auch dem höheren Körpergewicht zugeschrieben werden kann (Neafsey et al., 2009). Infektionen, Entzündungen und Krebs gehen mit einer erhöhten Zirkulation von Zytokinen, wie Interleukinen (IL1, IL6 und TNFα) einher. Diese Signalproteine triggern eine extreme Herunterregulierung der ADME-Genexpression in der Leber (Aitken et al., 2006; Klein et al., 2014). Einen wesentlichen Einfluss auf die Variabilität haben Interaktionen von weiteren Medikamenten (*drug-drug interactions*; DDI), die Nahrung und der Konsum von Drogen wie Alkohol oder Nikotin. Zum Beispiel induziert Ethanol die CYP2E1 Aktivität durch eine mRNA und Proteinstabilisierung (Ingelman-Sundberg et al., 1993). Eine Aktivierung kann ebenfalls über eine Induktion der Expression stattfinden, wie zum Beispiel die induzierende Wirkung von Rifampizin über PXR auf die CYP3A4 Expression (Staudinger and Lichti, 2008). Neben der Aktivierung ist die Inhibierung des ADME-Systems für einen Großteil der DDI verantwortlich. Die von Substraten vermittelte Inhibierung kann reversibel (kompetitiv oder nicht-kompetitiv) oder irreversibel sein. In beiden Fällen steigen die Spiegel des Substrats oder Metaboliten an und es können Nebenwirkungen oder toxische Zustände auftreten (Doligalski et al., 2012). Genetische Faktoren beeinflussen ebenfalls den Arzneimittelmetabolismus und werden im folgenden Abschnitt behandelt.

1.3 Pharmakogenetik

Der Begriff Pharmakogenetik beschreibt die Untersuchung von genetischen Faktoren, die den Metabolismus von Arzneimitteln und Xenobiotika beeinflussen können. In Zukunft soll durch dieses Wissen eine Vorhersage zur individuellen Reaktion auf ein Medikament und eine individuelle Anpassung der Dosierung möglich sein. So wird für jeden Patienten der therapeutische Nutzen optimiert und auftretende Nebenwirkungen

und Toxizität minimiert. Der systematische Ansatz einer genomweiten Suche nach Varianten, die einen Einfluss auf die Wirkung oder das Verhalten eines Medikaments zeigen, wird unter dem Begriff Pharmakogenomik zusammengefasst.

1.4 Genetische Variabilität– Ein allgemeiner Überblick

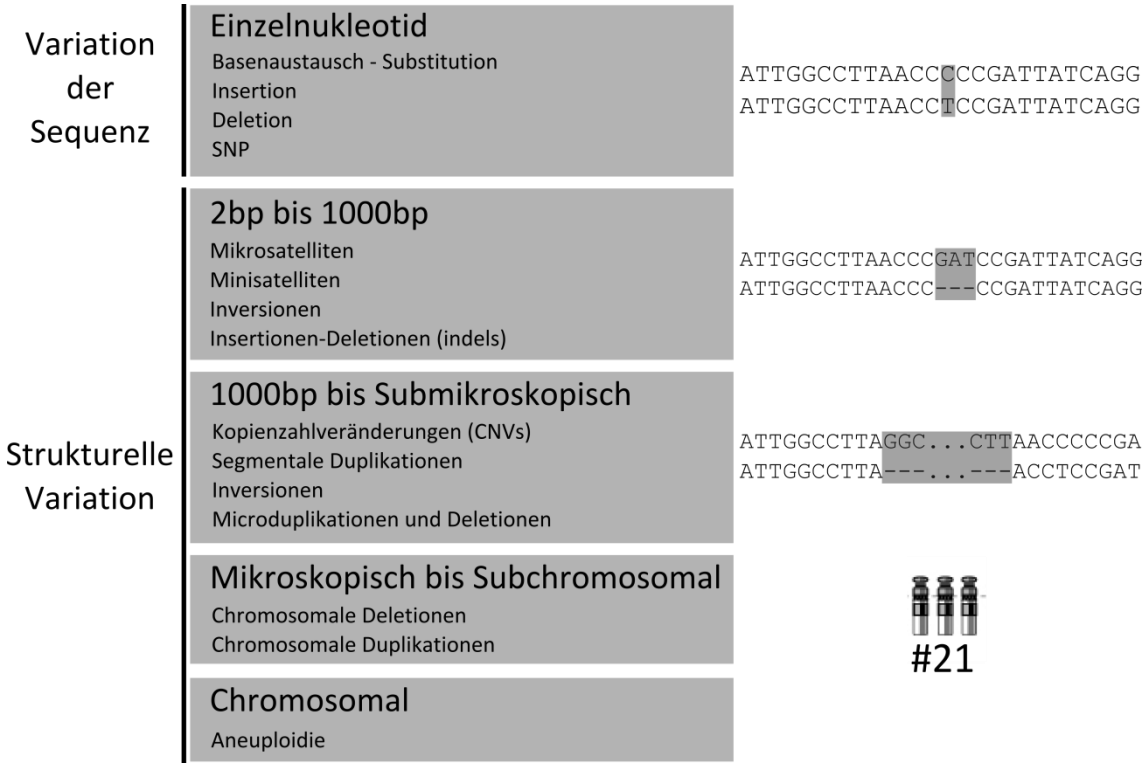


Abbildung 1: Übersicht der vorkommenden genetischen Variationen. Die Einteilung der strukturellen Varianten nach Größe erscheint willkürlich und basiert auf der historischen Auflösung von DNA Analysemethoden. Zytogenetische Verfahren wurden durch immer besser werdende molekulare Detektionsmethoden ersetzt. Adaptiert von Frazer et al. (2009) und Scherer et al. (2007).

Varianten des humanen Genoms werden in einer humanen Population als selten (<1%) oder häufig vorkommende Varianten definiert. Lässt sich eine Variante mit einer Häufigkeit von mehr als einem Prozent beobachten, wird der Begriff Polymorphismus verwendet. Eine weitere Kategorisierung erfolgt nach Anzahl der betroffenen DNA Nukleotide in Einzelbasenvarianten oder in strukturelle Varianten (Abbildung 1). Es finden sich weit mehr Einzelbasenmutationen (>99,9% von 4,1 bis 5 Millionen genetischen Varianten pro Person) als strukturelle Varianten im Genom einer einzelnen Person. Allerdings ist die absolute Anzahl der betroffenen Nukleotide durch strukturelle Varianten höher (1000 Genomes Project Consortium et al., 2015; Frazer et al., 2009). Einzelbasenpolymorphismen (*single nucleotide polymorphism*; SNP) können, je nach Lage im Genom und der Position in kodierenden Sequenzen, einen Einfluss auf die Expres-

sion und den Phänotyp nehmen. Funktionelle SNPs in regulatorischen Regionen (Promoter und Enhancerregionen) oder den Genen von Transkriptionsfaktoren beeinflussen die Transkription (cis-SNPs und trans-SNPs). Befindet sich ein SNP in einem kodierenden Bereich kann durch einen Aminosäureaustauch die Primärstruktur des Proteins abgeändert sein, wodurch die Proteinexpression oder die Aktivität beeinflusst wird (nicht-synonyme Varianten; Sadee et al., 2011). Aber auch synonyme SNPs und Varianten außerhalb der kodierenden Bereiche können sich auf die mRNA und Proteinexpression auswirken. Mechanismen, wie eine Destabilisierung der mRNA Struktur oder eine veränderte Regulation der Translation über ein alternatives Spleißen und Modulation der 3'UTR bindenden Faktoren, wie miRNAs sind vorstellbar. Für zahlreiche ADME-Gene und insbesondere für CYPs wurden SNPs identifiziert, die die Expression und Aktivität und damit die Wirkung von Medikamenten beeinflussen. Mutationen, die zu nicht-funktionellen Allelen und einer fehlender Proteinexpression führen, sind unter anderem für *CYP2A6* (*2; die nicht-synonyme Variante führt zu einem instabilen Protein, welches kein Häm binden kann), *CYP2D6* (*4; durch alternatives Spleißen bedingtes vorzeitiges Stopcodon) oder *CYP2C19* (*2; alternatives Spleißen) beschrieben. Eine umfangreichere Übersicht von funktionellen SNPs, speziell in Genen der CYP-Familie, findet sich in der Publikation Zanger and Schwab (2013).

1.4.1 Kopienzahlvariationen (CNVs)

1.4.1.1 Definition, Vorkommen, Entstehung und Detektion von CNVs

Kopienzahlvariationen (CNVs) stellen DNA Abschnitte dar, die im Vergleich zu einer Referenz dupliziert oder deletiert und laut Definition mindestens 1000bp lang sind. Diese Beschränkung erscheint etwas willkürlich gewählt und basiert auf dem schlechten Auflösungsvermögen älterer Detektionsmethoden. Allerdings ist der Vorteil dieser Einschränkung die Abgrenzung zu repetitiver DNA, wie Mikrosatelliten oder SINEs und LINEs. Die erste genomweite Studie zum CNV-Vorkommen benutzte einen BAC-Klon (*bacterial artificial chromosome*) Mikrochip mit einer Auflösung von 1Mb und identifizierte in 39 gesunden Personen insgesamt 255 CNV-Loci (Iafrate et al., 2004). Die Detektionsmethoden wurden kontinuierlich verbessert und 2006 beobachtete eine Arbeit mit 270 HapMap-Proben und einem Oligonukleotid-Mikrochip (Oligo-Mikrochip) 1447 CNV-Regionen, die ca. 12% des humanen Genoms abdeckten (Redon et al., 2006). Der Mittelwert der Variantenlängen wurde dabei im Vergleich von 341kbp auf 206kbp und die durchschnittlichen variablen DNA-Sequenzen im individuellen Genom von 24Mb auf 5Mb gesenkt. Ab diesem Zeitpunkt wurden hochauflösendere und präzisere Verfahren für eine CNV-Detektion, wie SNP-Mikrochips oder Sequenzierungen

der nächsten Generation (NGS) verwendet (Erläuterungen dazu siehe unten). Die Datenbank für genomische strukturelle Varianten (MacDonald et al., 2014) sammelt seither alle CNV-Informationen genomweiter Studien in gesunden Personen, stellt die Daten online frei zur Verfügung und erlaubt so eine Übersicht des variablen humanen Genoms (siehe1.5).

Es sind bisher vier Mechanismen diskutiert, die zur CNV-Entstehung beitragen. Dazu gehören die Retrotransposition, die nicht-allelische homologe Rekombination (NAHR), die Verbindung von nichthomologen Enden (non-homologeous end joining; NHEJ) und das Blockieren der Replikationsgabel und der Wechsel des DNA Strangs (fork stalling and template switching; FoSTeS) (Zhang et al., 2009). NAHR und NHEJ kommen bei DNA Doppelstrangbrüchen als Reparaturmechanismen in Frage, während ersteres auch in der Meiose zur Rekombination und damit zur CNV-Entstehung in der Keimbahn beiträgt. Während der NAHR lagern sich nicht-allelischen Sequenzen mit hoher Ähnlichkeit aneinander an und es kommt zum Austausch von Sequenzen. Eine Rekombination innerhalb eines Chromatids resultiert immer in einer Deletion. Eine Rekombination zwischen zwei Chromatiden führt zu einem reziproken Austausch oder einer Deletion und der reziproken Duplikation (Gu et al., 2008). Die Häufigkeit von NAHR korreliert mit dem Auftreten von Sequenzen mit hoher Sequenzsimilarität, deren Länge sowie deren Lage im Genom. Als Hotspots und Quellen sind Pseudo- und paraloge Gene sowie segmentale Duplikationen (SD) beschrieben (Sharp et al., 2005). Wie bereits angedeutet, kommen unterschiedliche Methoden zur Detektion von CNVs in Frage. Zu den genomweit anwendbaren Methoden gehören Oligo- und SNP-Mikrochips oder NGS-Verfahren. Die relative Kopienzahl von einzelnen Genen kann mit einer Echtzeit-PCR (qPCR) und speziellen TaqMan Assays bestimmt werden.

Oligo-Mikrochips basieren auf der Hybridisierung von zwei markierten Proben (Test und Referenz) an Sonden, typischerweise lange Oligonukleotide, die Teile des Genoms repräsentieren. Über den Vergleich der Signalintensitäten kann die Kopienzahl bestimmt werden. Der Einfluss der Referenzprobe auf das Ergebnis muss allerdings beachtet werden. Eine Deletion in der Referenzprobe ist dabei nicht von einer Duplikation in der Testprobe zu unterscheiden. Obwohl diese Chips neuerdings auf die eigenen Ansprüche und Bedürfnisse angepasst und mit hoher Sondendichte angeboten werden, wird heutzutage eher auf SNP-Mikrochips zurückgegriffen. Diese Chips bieten, neben der Genotypisierung von SNPs, eine Detektion von CNVs, die ebenfalls über den Vergleich der Signalintensitäten erfolgt (Redon et al., 2006). Ein Unterschied und der große Vorteil gegenüber Oligo-Mikrochips ist die Verwendung von log2 Intensitätsverhältnissen (log-R-Verhältnis; LRR) gegenüber allen Proben und nicht nur einer Referenzprobe. Zusätzlich können die jeweiligen berechneten Häufigkeiten der SNP-

Allele in einer Probe (B-Allelfrequenz; BAF) zur CNV-Detektion hinzugezogen werden (Wang et al., 2007). Durch dieses Vorgehen werden falsch-positive Ergebnisse minimiert. Durch die unterschiedliche Anzahl und Position der Sonden können sich Ergebnisse, die mit Mikrochips unterschiedlicher Hersteller gewonnen wurden, stark unterscheiden (Alkan et al., 2011). Eine weitere Methode ist ein NGS Ansatz, der in naher Zukunft die genetische Analyse mit Mikrochips ersetzen könnte (Wheeler et al., 2008). Die enorme Datenmenge und die dafür benötigte rechnerische und bioinformatische Leistung darf allerdings nicht außer Acht gelassen werden. Es gibt drei Strategien, die am häufigsten zur CNV-Detektion in NGS Daten verwendet werden. Alle analysieren das Anlagern der sequenzierten Segmente (*reads*) an das Referenzgenom und bestimmen über eine unnatürliche Signatur oder ein abweichendes Muster der reads die CNVs und deren Art (Alkan et al., 2011; Abbildung 2). Die Methode der gepaarten reads (*read pair*) vergleicht dabei die Spannweite und Orientierung der von beiden Seiten sequenzierten Segmente (paired-end) und erkennt Widersprüche gegenüber dem Referenzgenom. Die Bruchpunktkartierung (*split read*) untersucht den Beginn und das Ende eines CNVs und funktioniert am erfolgreichsten mit längeren reads (Medvedev et al., 2009). Die Analyse der Abdeckung (*read depth*) nimmt an, dass die Sequenzierung über alle Bereiche gleichförmig verläuft und die Anzahl der reads normalverteilt ist. Die Zahl der reads steigt also proportional mit der Kopienzahl einer Region an (Campbell et al., 2008). Dementsprechend werden in einer deletierten oder duplizierten Region weniger oder mehr reads zugeordnet (Medvedev et al., 2009; Abbildung 2).

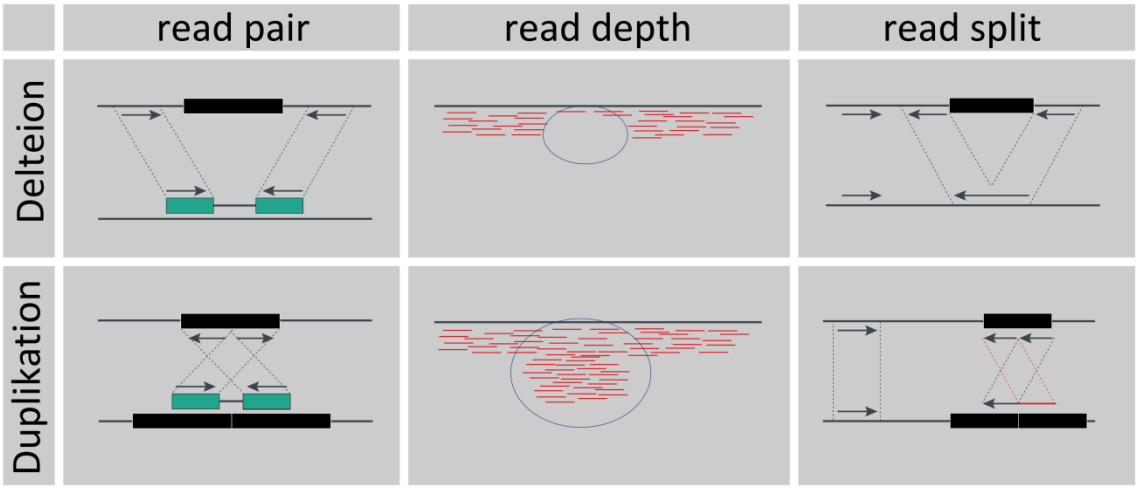


Abbildung 2: Übersicht über drei Ansätze mit denen man in NGS-Daten CNVs und deren Art (Deletion oder Duplikation) detektieren kann. Die Methode der gepaarten reads (*read pair*) analysiert die Zuordnung von gepaarten reads an das Referenzgenom. Die Abdeckungsanalyse (read depth) bestimmt über die Zu- oder Abnahme der reads die Kopienzahl. Die dritte Methode (split reads) detektiert während der Zuordnung an das Referenzgenom die exakten Bruchpunkte der Variation. Dieses Verfahren benötigt in der Regel längere reads als die anderen Methoden. Das Schema wurde von Alkan et al. (2011) übernommen.

1.4.2 Konsequenzen von strukturellen Varianten auf den Phänotyp

CNVs können evolutiv betrachtet als Spielwiese für eine Diversifikation und eine Spezialisierung von paralogen Genen dienen (Dumas et al., 2007). Betreffen CNVs eine vollständige Genregion inklusive regulatorischer Sequenzen (dann gilt ein Gen in dieser Arbeit als CNV-assoziiert), kann die Expression des Gens verändert sein. Ein erster Versuch, den Einfluss von CNVs genomweit in humanen lymphoblastoiden Zelllinien (LCLs) abzuschätzen, zeigte, dass CNVs 20% der Expressionsvariabilität erklären (Stranger et al., 2007). Obwohl weitere Studien mit Fliegen, Mäusen, Ratten und gesunden humanen Proben eine positive Korrelation zwischen Gendosis und Expression bestätigten, wurden auch CNV-assoziierte Gene gefunden, deren Expression negativ zur Kopienzahl korrelierte oder trotz Kopienzahl nicht verändert war (Guryev et al., 2008; Henrichsen et al., 2009a; Woodwark and Bateman, 2011; Zhou et al., 2011). Die Gründe für das Gendosis unabhängige oder reverse Verhalten sind vielfältig und noch nicht vollständig geklärt. In Frage kommende molekulare Mechanismen der Dosiskompensation umfassen unter anderem einen unvollständigen Einschluss der Promoter oder Enhancer Region im CNV-Segment, eine differentielle genomische Prägung, eine generelle monoallelische Expression des CNV-assoziierten Gens, eine Regulation über miRNAs, gelinkte genetische Varianten und negative Feedbackschleifen (Henrichsen et al., 2009b; Woodwark and Bateman, 2011). Im Gegensatz zu den Ergebnissen aus gesundem Gewebe zeigen in Krebsgewebe nahezu alle CNV-assoziierten Gene eine dosis-sensitive Expression, wenn man die Genregulation als Kovariate in die Analyse mitaufnimmt (Fehrmann et al., 2015). Neben dem direkten Einfluss sind auch cis- und trans- Effekte vorstellbar (Harewood et al., 2012). Der Einfluss von CNVs auf Krankheiten (z.B. Schizophrenie, Alzheimer-Krankheit, Autismus), auf das Immunsystem (z.B. α- und β-Defensin), auf eine Krebserkrankung, auf den Stoffwechsel (z.B. Adipositas) und den Arzneimittelmetabolismus sind bekannt (Beckmann et al., 2008; Eichelbaum et al., 2006; Friedrichsen et al., 2013; Machado and Ottolini, 2015).

1.4.3 Wichtige funktionelle CNVs von ADME-Genen

Wichtige pharmakogenetische CNVs wurden bereits in den 90er Jahren in der *CYP2D6* Genregion gefunden und deren Einfluss auf die Enzymaktivität gezeigt (Gaedigk et al., 1991; Johansson et al., 1993). Ein weiteres Gen der Phase I, in dessen Genregion CNVs auftreten, ist *CYP2A6*. Unter den Phase II Genen finden sich außerdem CNVs in *GSTM1*, *GSTT1*, *SUL1A1* und *UGT2B17* (Johansson and Ingelman-Sundberg, 2008).

Die Deletion von *CYP2D6* (*CYP2D6*5*) betrifft das komplette Gen. Individuen mit einer homozygoten Deletion weisen also keine Proteinexpression und damit keine Enzymaktivität auf und der Metabolismus von CYP2D6 Substraten ist ultimativ beeinträchtigt. Die betroffenen Personen werden als Langsam-Metabolisierer (*poor metabolizer*; PM) klassifiziert. Die *CYP2D6* Genedeletion *5 kommt weltweit in 3-5% der Individuen vor. Im Gegenzug wurde gezeigt, dass in Personen mit einer Duplikationen von funktionellen *CYP2D6* Allelen (*CYP2D6*1*, *CYP2D6*2 und CYP2D6*35*) die CYP2D6 Enzymaktivität gesteigert ist. Die Biotransformation ist somit in Individuen, die eine Duplikation erben, schneller. Sie werden deshalb als schnelle Metabolisierer (ultrarapid metabolizer; UM) bezeichnet (Zanger and Schwab, 2013). 1-5% der Personen europäischer Abstammung tragen ein dupliziertes *CYP2D6* Allel. In manchen Populationen afrikanischer Herkunft liegt die Frequenz bei 10-50% (Ingelman-Sundberg, 2005). Neben den Deletionen und Duplikationen kommen weitere Allele und Kombinationen mit SNPs (siehe oben) und Genkonversionen mit dem Pseudogen *CYP2D7P1* vor (Gaedigk et al., 2012; Schaeffeler et al., 2003). Die humane CYP-Allel-Nomenklatur-Datenbank (http://www.cypalleles.ki.se/; Zugriff August 2015) listet aktuell mehr als 75 Allele für den *CYP2D6* Genlocus auf. Für eine Vielzahl an Medikamenten wie Antiarrhythmika (Propafenon, Metoprolol), Antidepressiva und Opioide wie Codein, wurde ein genotypabhängiger Metabolismus festgestellt. Letzteres wird von CYP2D6 pharmakologisch aktiviert und der *CYP2D6* Genotyp beeinflusst somit die Wirkungseffizienz und die Sicherheit des Schmerzmittels. In Personen mit mehreren Genkopien (UMs) führt der erhöhte Metabolismus zu lebensbedrohlichen Opioid-Vergiftungen (Madadi et al., 2007).

Auch in der *CYP2A6* Genregion wurden sowohl Deletionen (*CYP2A6*4A-H)* als auch Duplikationen (1X2A and *1X2B) beobachtet, die beide jeweils die Expression und Enzymaktivität beeinflussen (Fukami et al., 2007; Rao et al., 2000). Neben den CNVs kommen weitere nicht-funktionelle Allele wie *9, *35 und Hybridallele mit dem Pesudogen *CYP2A7* (*12) vor (Haberl et al., 2005). Es wurde gezeigt, dass der Metabolismus von Nikotin durch die *CYP2A6* CNVs beeinflusst wird und dadurch das Rauchverhalten und die Suchtanfälligkeit in *CYP2A6* CNV-Trägern verändert ist (Mwenifumbo and Tyndale, 2007).

Duplikationen der kompletten *CYP2E1* Genregion (10q26.3) sind in Individuen mit europäischer, afrikanischer und asiatischer Abstammung beschrieben (MacDonald et al., 2014; Martis et al., 2013). Bisher zeigte eine Assoziationsstudie einen Einfluss der 10q26.3 Duplikation, die neben *CYP2E1* auch weitere Gene einschließt, auf Adipositas und Stoffwechselparameter (Yang et al., 2013). Allerdings konnte eine weitere Studie keinen Zusammenhang zwischen *CYP2E1* CNVs und der Reaktion auf Alkohol fest-

stellen (Webb et al., 2011). Der CNV-Einfluss auf die CYP2E1 Expression und die Enzymaktivität wurde bisher noch nicht molekular entschlüsselt.

Während Deletionen von *GSTM1* und *GSTT1* sehr häufig sind (in 50% und 30% der Individuen mit europäischer Abstammung), fehlen Duplikationen weitestgehend in den meisten Populationen (He et al., 2011). Beide Enzyme sind, wie bereits erwähnt, in der Detoxifizierung von endogenen und exogenen Substanzen beteiligt. Die jeweilige homozygote Deletion beider Gene ist deshalb mit unterschiedlichen Krebsentitäten, wie Kolonkrebs, Leukämie und diversen Toxizitäten assoziiert (Das et al., 2009).

Pharmakogenetische Studien zeigten, dass *SULT1A1* ebenfalls von CNVs betroffen ist. So sind Deletionen und häufige Duplikationen des Gens beschrieben (Hebbring et al., 2007). Wobei Deletionen bisher nur in Personen mit europäischer Abstammung (4.7%), nicht aber in Individuen mit afrikanischer Abstammung beschrieben sind. Dagegen finden sich drei oder mehr Kopien in 26% bzw. 62% der Individuen (Hebbring et al., 2009). Die CNVs korrelieren signifikant mit der Enzymaktivität und erklären einen großen Teil der Aktivitätsvariabilität (Hebbring et al., 2009; Yu et al., 2010).

1.5 Die Datenbank für genomische strukturelle Varianten

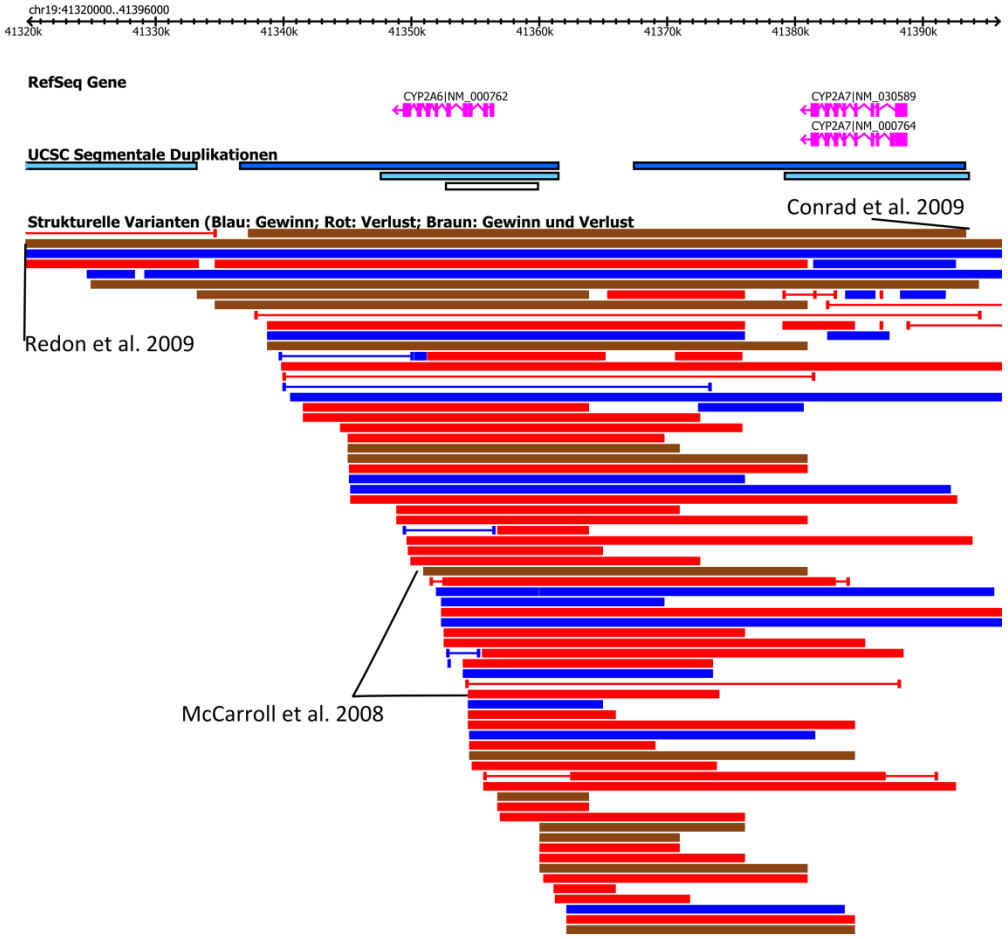


Abbildung 3: DGV-Ausschnitt des Chromosoms 19 (19q13.2) mit der *CYP2A6-CYP2A7* Genregion im Zusammenhang mit lokalen genomischen Eigenschaften. Von oben nach unten: Position von bekannten Genen der NCBI Referenzsequenzkollektion. Segmentale Duplikationen (von UCSC), die als Hotspots der CNV-Entstehung über NAHR diskutiert werden. Die unterschiedliche Einfärbung zeigt die Sequenzsimilarität. In Dunkelblau sind Sequenzen mit 96-98,9% Ähnlichkeit, in Hellblau sind Sequenzen mit 93-95,9% Ähnlichkeit, in Weiß sind Sequenzen mit 90-92,9% Ähnlichkeit dargestellt. Die in DGV hinterlegten strukturellen Varianten (Zugriff August 2015). Jeder Balken entspricht einem identifizierten CNV-Segment in einer Studie. Blaue Balken beschreiben einen Sequenzgewinn, rote Balken einen Verlust und braune Balken beschreiben ein Segment, welches sowohl als dupliziert als auch deletiert detektiert wurde. Dünne Linien zeigen Varianten, deren Bruchpunkte unsicher und häufig ein Ergebnis aus Studien mit geringem Auflösungsvermögen sind.

Die Online-Datenbank für genomische strukturelle Varianten (DGV) sammelt CNV-Informationen von wissenschaftlichen Publikationen, die genomweit strukturelle Varianten in gesunden Personen identifiziert haben. Eingeschlossen werden dabei nur Varianten, die größer als 50bp und kleiner als 3Mb sind (MacDonald et al., 2014). Die in der Datenbank gelisteten Studien haben eines der oben erwähnten Verfahren zur CNV-Detektion verwendet, wobei Oligo- und SNP-Mikrochips die meiste Anwendung

fanden. Wie bereits beschrieben, beeinflussen Technologie und Verfahren der CNV-Bestimmung die Länge und Position der detektierten strukturellen Variante. Ein BAC-Klon Mikrochip tendiert dazu, die Grenzen einer Variante zu überschätzen. Im Gegensatz dazu eignen sich NGS-Methoden besser zur Detektion von kürzeren Varianten mit einer medianen Länge von 740bp. Allerdings treten Schwierigkeiten bei der Identifizierung von sehr langen CNV-Segmenten auf (Alkan et al., 2011). Wie in Abbildung 3 zu erkennen ist, wurden im *CYP2A6* Genlocus von einer Vielzahl von Studien CNVs mit unterschiedlicher Größe und Position detektiert. Dabei kann sich die Lage und Position von Segmenten in ein und derselben Probe unterscheiden. Dies ist exemplarisch für die Studien von Redon et al. (2006), McCarroll et al. (2008) und Conrad et al. (2009) gezeigt (Abbildung 3). Die drei Studien verwendeten zwar dieselben HapMap-Proben, aber unterschiedliche Verfahren mit unterschiedlichem Auflösungsvermögen. Darüber hinaus sagt die Anzahl der annotierten Segmente einer Genregion nichts über die Frequenz der in der Region vorkommenden CNVs aus. In DGV sind CNVs einer Studie, die in mehreren Proben an ungefähr der gleichen Position und mit der gleichen Länge detektiert wurden, zu einer CNV-Region (CNVR) zusammengefasst. Aufgrund dieser Einschränkungen sollte bei der Auswertung der Datenbank folgende Faustregel befolgt werden: DNA-Regionen, die in unterschiedlichen Studien oder mit verschiedenen Methoden als CNV identifiziert wurden, sind verlässlichere und wahrscheinlichere CNV-Kandidaten als Regionen, in denen nur Einzelbeobachtungen oder nur sehr lange Varianten detektiert wurden. Um als CNV-assoziierter Kandidat zu gelten, sollte deshalb ein Gen von mindesten zwei Varianten, die von unterschiedlichen Studien detektiert wurden, komplett überlappt werden (MacDonald et al., 2014).

1.6 Ziele dieser Arbeit

Da es zu Beginn dieser Arbeit keine umfassende Untersuchung zum Vorkommen von strukturellen Genvarianten und der funktionellen Bedeutung von Kopienzahlvarianten bei ADME-Genen gab, sollte zunächst das Vorkommen solcher Varianten in den wichtigsten Genen des Arzneimittelmetabolismus und -transports (ADME-Gene) systematisch untersucht und dann der funktionelle Einfluss ausgewählter Beispiele auf den Phänotyp analysiert werden. Im Einzelnen umfasste dies folgende Schritte:

- Analyse der öffentlichen Datenbank für genomische strukturelle Varianten (DGV) zum Vorkommen von CNVs in ADME-Genen.
- Bestimmung der Kopienzahl aller ADME-Gene in humanen Leberproben der Kohorte IKP148. Validierung bereits bekannter Gene mit Kopienzahlvarianten (zB *CYP2A6*, *CYP2D6*, *GSTM1*) und deren Funktionalität anhand quantitativer Analysen.
- Entwicklung einer Methode zur Nutzung der neu vorhandenen NGS-Daten und deren Anwendung. Validierung der NGS-Ergebnisse für ausgewählte Gene mit Hilfe von quantitativer PCR unter Verwendung spezifischer TaqMan CNV-Assays.
- Detaillierte funktionelle Untersuchung besonders interessanter Kandidatengene. Hierzu sollten weitere Phänotypen wie Proteinwerte und Enzymaktivitätsdaten erhoben und statistisch ausgewertet werden.

In anschließenden Analysen sollte weiterhin untersucht werden, ob CNVs auch mit SNPs in Zusammenhang stehen, um so möglicherweise eine einfachere Genotypisierung zu ermöglichen. Schließlich sollte ein Vergleich zwischen CNVs und SNPs hinsichtlich ihrer Effekte auf die Funktionalität der ADME-Gene wichtige Basisdaten für die Pharmakogenetik liefern.

2 Material und Methoden

2.1 Material

2.1.1 Chemikalien

Die verwendeten Chemikalien wurden von folgenden Herstellern bezogen. Bio-Rad (München, Deutschland), Merck (Darmstadt, Deutschland), Sigma-Aldrich (Steinheim, Deutschland), Roth (Karlsruhe, Deutschland), Biozym (Oldenburg, Germany), Invitrogen (Karlsruhe, Deutschland), Knoll (Ludwigshafen, Deutschland), Ambion (Ambion, Applera GmbH), GIBCO (Carlsbad, USA), Promega (Madison, USA).

2.1.2 Puffer und Lösungen

Tabelle 1: Zusammensetzung von Puffern und anderen Lösungen

Bezeichnung	**Bestandteile**	**Menge**
APS 10%	Ammoniumperoxodisulfat	1g
	H2OMillipore	ad 10ml
Elektrophorese Puffer (10x)	Tris-HCl	150g
	Glycin	720g
	SDS (20%)	250ml
	H2OMillipore	ad 5000ml
Lämmli Auftragspuffer (5x)	TRIS HCl pH 6,8	60mM
	β-Mercaptoethanol	14,4mM
	SDS	2%
	Glycerin	24%
	Bromphenolblau	1%
Magermilch/TBST 5%	Magermilchpulver	10g
	1x TBST	ad 200ml
TBS(10x)	TRIS	150g
	NaCl	400g
	KCl	10g

Bezeichnung	Bestandteile	Menge
	H2OMillipore	ad 5000ml
	HCl (pH7,4)	
TPST (1x)	10x TBS	500ml
	50% Tween 20	10ml
	H2OMillipore	ad 5000ml
Transferpuffer (1x)	TRIS-HCl	9g
	Glycin	14,6g
	SDS (10%)	18,5ml
	Methanol	1000ml
	H2OMillipore	ad 5000ml
TRIS-HCl 0,5 M, pH 6,8/pH 8,8	TRIS	30g
	H2OMillipore	ad 500ml
	HCl (pH6,8 oder 8,8)	

2.1.3 Geräte und Laborbedarf

Tabelle 2: Verwendete Geräte, Materialien und Laborbedarf

Gerät oder Material	Hersteller
96-Mikrotiter Platte (mit Barcode)	Applied Biosystems, Darmstadt, Deutschland
Nitrozellulose Membran	Whatman, Dassel, Deutschland
Millipore Anlage	Millipore, Molsheim, Frankreich
Nanodrop 2000c	Peqlab Biotechnologies GmbH, Erlangen, Deutschland
pH Meter	Schott, Mainz, Deutschland
Protein-Molekulargewichtsmarker (Precision Plus)	Bio-Rad, München, Deutschland
Vortexer	BIOZYM Scientific GmbH, Oldendorf
7900 HT Real Time PCR System (TaqMan)	Applied Biosystems, Darmstadt, Deutschland
ODYSSEY Infrared Imaging System	LI-COR Biosciences GmbH, Bad Homburg, Deutschland
Heizblock CLF	Laborgeräte, Emersacker, Deutschland
Fastblot-Apparatur	Biometra, Göttingen, Deutschland

Gerät oder Material	Hersteller
SDS-PAGE Kammer	Bio-Rad, München, Deutschland
Fuji LAS-1000 CCD Kamera	Raytest, Straubenhardt, Deutschland
Zentrifuge 5414 C	Eppendorf AG, Hamburg, Deutschland

2.1.4 Antikörper

Tabelle 3: Weitere Materialien und Laborbedarf.

Antikörper	Hersteller, Klon	Referenz
Anti-CYP2E1 (Maus, monoklonal, 1:10000)	2-106-12/ Frank Gonzalez	(Gelboin et al., 1996)
Anti-SULT1A1 (Kaninchen, Serum, 1:10000)	Hans Ruedi Glatt	-
Anti-Maus IRD800 (Ziege, 1:10000)	LI-COR Biotechnology, Lincoln, USA	
Anti- Kaninchen IRD800 (Ziege, 1:10000)	LI-COR Biotechnology, Lincoln, USA	
Anti-Maus Peroxidase (Ziege, 1:10000)	DC02L, Oncogene	

2.1.5 20x TaqMan Assays®

2.1.5.1 mRNA Expressionsassays

Zielgen	Assay Nummer/ Referenz	Zielbereich
CYP2D6[1]	(Toscano et al., 2006)	Exons 5-7
CYP2A6[1]	(Haberl et al., 2005)	Exons 5-6
CYP2E1	Hs00559368_m1	Exons 6-7
RPLP0	4326314E	Exon 3

[1]Selbst designed

2.1.5.2 Assays für die Kopienzahlbestimmung

Zielgen	Assay Nummer	Zielbreich
CYP2D6	Hs00010001_cn	Exon 9
CYP2A6	Hs07545275_cn	Intron 7
CYP2E1	Hs00010003_cn	Promoter
CYP2E1	Primer 1:	Exon 5

Zielgen	Assay Nummer	Zielbreich
	5’-ACACACACAATTATGTTGCATCCA-3’	
	Primer 2:	
	5’-ACATACTCTTTTACTTCAGCCACATTTT-3’	
	Sonde:	
	6-FAM-AGCTTTCTACACTACTTGC-MGB	
SULT1A1	Hs04461762_cn	Promoter
POR	Hs03624286_cn	Intron 2

2.1.6 Verwendete Software,R-Pakete und Webseiten

Name	Quelle	Referenz
Analyst Software v9.0	www.genedata.com	Genedata, Basel, CH
biomaRt_2.20.0	http://www.biomart.org/	(Durinck et al., 2009)
BLAST	http://goo.gl/MMvd6b	NCBI
cn.mops_1.12.0	cn.mops	(Klambauer et al., 2012)
CopyCaller v2.0	www.copycaller.html	Applied Biosytems
DGV	http://dgv.tcag.ca	(MacDonald et al., 2014)
Ensembl	http://www.ensembl.org	-
GenomicRanges_1.18.4	GenomicRanges	(Lawrence et al., 2013)
GOrilla	cbl-gorilla.cs.technion.ac.il	(Eden et al., 2009)
Gviz_1.10.10	Gviz	-
Haploreg	www.haploreg.php	(Ward and Kellis, 2011)
LALIGN	www.LALIGN.html	(Homer et al., 2009)
Linux	http://www.ubuntu.com	-
MASS_7.3-33	MASS	(Venables & Ripley 2002)
MicroSnipers	www.microsniper.de	-
NCBI	http://www.ncbi.nlm.nih.gov/	-
Office 2010	https://www.microsoft.com	-
PLINK	plink	(Purcell et al., 2007)
PolymiRTS 3.0	compbio.uthsc.edu	-
Proteinatlas	http://www.proteinatlas.org	(Uhlén et al., 2015)
qqman_0.1.2	http://goo.gl/NvAMz6	(Turner, 2014)
quantreg-5.05	quantreg	(Koenker, R., 2015)

Name	Quelle	Referenz
REVIGO	revigo.irb.hr	(Supek et al., 2011)
SNPassoc_1.9-2	https://goo.gl/QtHWva	-
Software R v3.1.2	http://cran.r-project.org/	(R Core Team, 2014)
UCSC	https://genome.ucsc.edu/	(Kent et al., 2002)

2.2 Methoden

2.2.1 Vergleich der in DGV gelisteten CNVs mit den Positionen der ADME-Genloci

Die Information über die CNVs umfasste die chromosomale Lage, die verwendete Methode und die Probenherkunft und wurden von DGV bezogen (Zugriff am 16.10.2014; http://dgv.tcag.ca/dgv/docs/ GRCh37_hg19_supportingvariants_2014-10-16.txt). Die in DGV hinterlegten Studien wurden nach verwendeter Probenanzahl (n> 40), Methode (keine BAC Klone) und den insgesamt detektierten CNVs (n>= 250) gefiltert. Der Start und das Ende der CNVs und ADME-Gene wurden mit dem „GenomicRanges“-Paket 1.16.4 (Lawrence et al., 2013) und der Software R 3.1.2 (R Core Team, 2014) verglichen. Es wurden nur CNVs aufgenommen, die das untersuchte Gen komplett einschlossen und von mindestens zwei verschiedenen Studien detektiert wurden. Die angegebene Frequenz stellt eine Genotypen-Frequenz dar und wurde wie folgt kalkuliert: Studienweise wurde jeweils die Anzahl der Proben, in denen die Variante gefunden wurde, durch die maximale Probenzahl geteilt. Anschließend wurden studienübergreifend Mittelwerte berechnet.

2.2.2 Die HapMap-Proben

2.2.2.1 Analyse der CNVs in ADME-Genloci in HapMap-Proben

Die populationsabhängige Analyse der CNV-Verteilung in ADME-Genen basiert auf den obengenannten Daten (siehe 2.2.1). Der Datensatz wurde allerdings auf HapMap-Proben, die aus dem asiatischen (CHB= Han Chinese aus Beijing, China; n= 112; JPT= Japaner aus Tokyo, Japan; n= 105), dem afrikanischen (YRI= Yoruba aus Ibadan, Nigeria; n= 195) und dem europäischen (CEU= Einwohner mit Nord- und Westeuropäischer Vorfahren aus Utah, USA; n= 184) Raum stammten, reduziert. 96% der Proben wurden von unabhängigen Studien mehrmals zur CNV-Detektion verwendet (Tabelle A2). Wurden unterschiedliche Ergebnisse (z.B. Studie A detektiert eine

Deletion, Studie B eine Duplikation in Probe 1 und Gen X) festgestellt, wurden die Resultate genspezifisch angepasst oder aus der weiteren Analyse genommen.

2.2.2.2 mRNA Expressionsprofile

Populationsunabhängig normalisierte mRNA Expressionsdaten waren online frei zugänglich (Stranger et al., 2007). Die Daten wurden in lymphoblastoiden Zelllinien (LCL) von 270 HapMap-Proben mit dem Illumina WG6 Expressionsmikrochip erhoben. Die exakte Detektion niedrig exprimierter Gene ist bei dieser älteren Chipgeneration problematisch, da eine Korrelation zwischen detektierbarer Expressionsvariation und Expressionsstärke beobachtet wurde (Schuster-Böckler et al., 2010; Abbildung 4). Deshalb wurden 33.144 von 47.297 Sonden mit einer Standardabweichung kleiner sieben von der folgenden Analyse ausgeschlossen, darunter auch Sonden für ADME-Gene (Abbildung 4; graue und rote Punkte). Insgesamt standen so für 13.173 Gene mRNA Expressionen für die Assoziationsanalyse zur Verfügung.

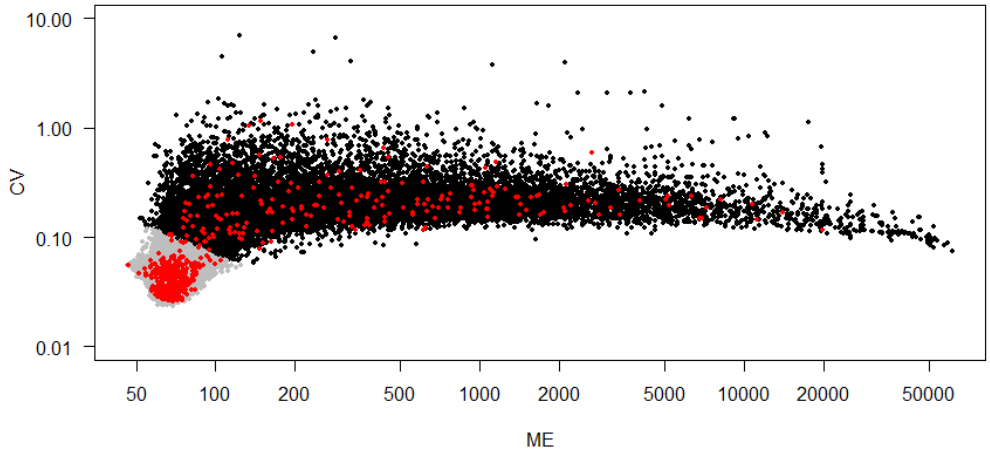


Abbildung 4: Dargestellt ist der Variationskoeffizient (CV) über der durchschnittlichen Expressionsstärke (ME). Die Variation der Genexpression hängt mit der Sensitivität der verwendeten Mikrochip Technologie zusammen. Expressionen mit einer Standardabweichung kleiner sieben wurden, wie in Schuster-Böckler et al. (2010) beschrieben, entfernt (grau eingefärbt). In Rot sind die in dieser Arbeit betrachteten 340 ADME-Gensonden markiert.

2.2.2.3 Assoziationsanalyse zwischen CNVs und der mRNA Expression in HapMap-Proben

Die Korrelation zwischen Kopienzahl und mRNA wurde mit einer Varianzanalyse (ANOVA) für eine multivariate Medianregression berechnet. Für jede Genexpression wurden zwei Medianregressionsmodelle erstellt, mit der Funktion anova.rq des R-Pakets quantreg_5.11 (Koenker, R., 2015) mit den Parametern Rang Test-Statistik und „Wilcoxon-score-Funktion“ miteinander verglichen und der p-Wert aus der Analyse no-

tiert. Dabei umfasste ein Modell die unabhängige Variable der Herkunft der Proben (CEU, YRI oder CHB&JPT), das zweite Modell enthielt zusätzlich die Kopienzahl. Die Kopienzahl war dabei wie in DGV üblich als Verlust (kleiner zwei Genkopien), Normal (zwei Genkopien) und als Gewinn (mehr als zwei Genkopien) kodiert.

2.2.3 TCGA-Lebergewebedaten

In der Datenbank des TCGA-Projektes existieren frei zugängliche Datensätze von gepaarten Biopsien aus einem hepatozellulären Karzinom und gesundem Lebergewebe. Darunter befinden sich prozessierte CNV-Daten und mRNA Expressionswerte zu den jeweiligen Geweben, sowie klinische und demografische Informationen zu den Donoren. Alle verwendeten Daten wurden vom TCGA-Forschungsnetzwerk erhoben und prozessiert (http://cancergenome.nih.gov/) und als komprimierte Matrix über das TCGA-Downloadportal bezogen (tcga-data.nci.nih.gov; Zugriff am 02.05.2014). Zum Zeitpunkt dieser Arbeit bestand die Kohorte aus 351 Patienten für die Daten aus Tumor und teilweise gepaarten gesunden Gewebeproben vorlagen (Tumor, n=351; Blut, n=269; histologisch normales Lebergewebe, n= 82). Eine Übersicht über die nichtgenetischen Faktoren liefert folgende Tabelle 4.

Tabelle 4: Demographie und klinische Dokumentation der Leberspender des TCGA-Projektes.

Gruppe	Anzahl[1]
Geschlecht	
weiblich	105
männlich	228
Altersangabe	
< 50	69
≥ 50	264
Abstammung	
Afrikanisch	17
Asiatisch	151
Europäisch	156
Histologische Diagnose	
Hepatozelluläres Karzinom	326
Cholangiokarzinom	7

[1]Die Summe in den einzelnen Gruppen beträgt wegen fehlender Informationen nicht immer 351.

2.2.3.1 CNV-Daten der TCGA-Leberproben

Das TCGA-Konsortium stellt prozessierte Intensitätsdaten (Level 3) eines SNP-Mikrochips (Affymetrix 6.0) zur Verfügung. Dabei wurden vom Konsortium bereits Mikrochipsonden mit gleicher Intensität in Segmente zusammengefasst und ein Segment-Mittelwert (sm) mit einem zirkulär-binären Segmentierungsalgorithmus wie folgt kalkuliert: Die sm-Werte wurden aus dem Mittelwert des logarithmierten Verhältnisses der Intensität gegenüber einer Referenz aller Proben im Segment [log2(gemessene Intensität/Referenzintensität)] berechnet. Sie geben den Hinweis, ob das Segment normal, dupliziert oder deletiert vorliegt.

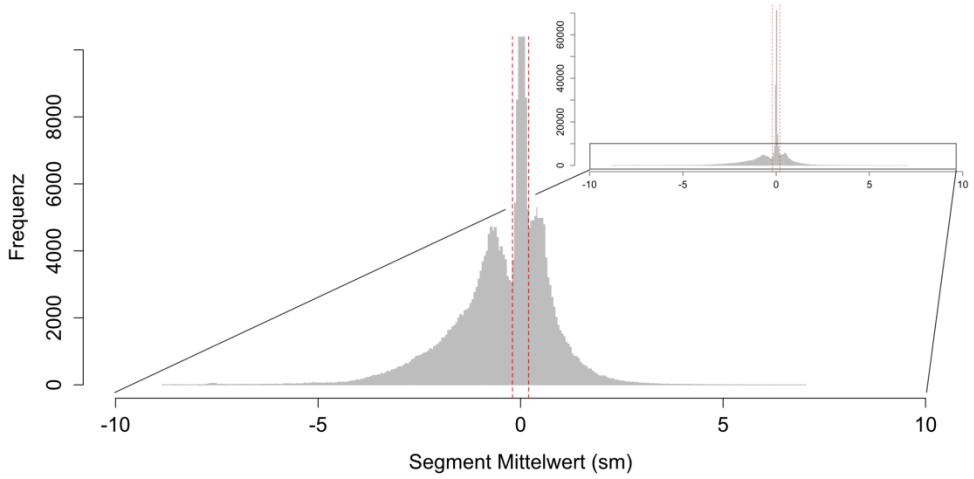


Abbildung 5: Histogramm der CNV-Segment-Mittelwerte (sm) aller 760 Gewebeproben (Tumor, Blut und normales Gewebe). Die verwendeten Grenzwerte sind als vertikale Linien in Rot eingezeichnet. Bei einem Wert kleiner als -0,2 wurde von einer Deletion ausgegangen. Ein größerer Wert als 0,2 wurde als Amplifikation des Segmentes definiert.

Ein Grenzwert, ab dem ein Segment als deletiert oder dupliziert vorliegt, wurde mit Hilfe einer Verteilung aller sm-Werte bestimmt (Abbildung 5 und Kim et al., 2015). Segmente mit einem Wert sm≤ -0,2 wurden als Deletion mit der Kopienzahl eins festgelegt. Als Duplikation und der Kopienzahl drei wurden Segmente mit einem Wert sm≥ 0,2 definiert (Li et al., 2013). Um falsch positive Ergebnisse zu minimieren, wurden Segmente, die aus weniger als 10 Sonden zusammengesetzt waren, ausgeschlossen. Drei Ausreißerproben mussten im Verlauf der Analyse wegen einer zu hohen CNV-Detektionsrate ausgeschlossen werden, was zu einer finalen Kohorte von 348 Proben führte. Um die Verlässlichkeit der Methode und verwendeten Grenzwerte einschätzen zu können, wurden die strukturellen Mutationen in der Keimbahn mit CNV-Informationen der DGV-Datenbank verglichen. Dabei musste mindestens 50% des CNV-Bereichs mit einer publizierten CNV übereinstimmen, um als bekannt gewertet zu

werden. CNVs in den 340 ADME-Genloci wurden, wie in 2.2.1 bereits beschrieben, extrahiert. Alle Positionsangaben bezogen sich auf das Referenzgenom GRCh37 (hg19).

2.2.3.2 Auswertung der Genexpressionsdaten einer RNA Sequenzierung

Normalisierte Level 3 „Read-Count“ Daten waren für 50 gesunde Lebergewebeproben und für 343 Tumorproben verfügbar. Der Read-Count beschreibt die Anzahl der Sequenzierungen die jedem Gen erfolgreich zugeordnet werden konnten. Expressionsunterschiede zwischen gesundem und Tumorgewebe und zwischen Proben unterschiedlicher Abstammung (siehe Tabelle 4) wurden mit dem R Paket edgeR_3.8.6 (Robinson et al., 2010) und der generalisierten linearen Modellfunktion mit Standardeinstellungen untersucht. Die anschließende Gen-Ontologie (GO) Analyse wurde mit den Online-Werkzeugen Gorilla (cbl-gorilla.cs.technion.ac.il) und Revigo (revigo.irb.hr) vorgenommen.

2.2.3.3 Analyse der Assoziation von CNVs und mRNA Expression im TCGA-Datensatz

Die Korrelationsanalyse von CNVs und den normalisierten RNA-Seq Genexpressionendaten (rpkm) wurde, wie bereits in 2.2.2.3 beschrieben, mit einer Varianzanalyse für lineare mediane Regressionsmodelle durchgeführt.

2.2.4 Die Leberproben der Studie IKP148

Proben von Blut und Lebergewebe wurden im Rahmen einer prospektiven Studie von 1999 bis 2003 in einer Biobank des Dr. Margarete Fischer-Bosch-Instituts für Klinische Pharmakologie, Stuttgart gesammelt. Die Leberchirugischen Gewebeentnahmen erfolgten am Chirurgiezentrum der Charité (Campus Virchow-Klinikum, Humboldt Universität, Berlin). Die Studie wurde von der Ethikkommission der Medizinischen Fakultät der Charité und der Universität Tübingen genehmigt, sowie nach den Statuten der Deklaration von Helsinki durchgeführt. Die Patienten willigten nach Aufklärung schriftlich ein. Das Gewebe wurde von einem Pathologen überprüft und nur histologisch tumorfreies Gewebe wurde aufgenommen und bei -80°C gelagert. Die klinische und demographische Dokumentation umfasste unter anderem das Alter, das Geschlecht, die Diagnose, die zur Biopsie führte, Medikation vor der Operation, Leberfunktionsparameter wie dem CRP-Spiegel, sowie eine Selbsteinschätzung zu Alkohol und Nikotinkonsum. Patienten mit Hepatitis, Zirrhose oder chronischem Alkoholkonsum wurden ausgeschlossen. Insgesamt konnte so auf 150 Leberproben zurückgegriffen werden. Eine

Übersicht findet sich in Klein et al., (2010). Genomische DNA lag bereits isoliert vor (Wolbold et al., 2003).

2.2.4.1 Identifizierung von Kopienzahlveränderungen in einem ADME-Exonsequenzierungspanel

In Zusammenarbeit mit der Diagnostik und Sequenzierfirma CeGaT (Tübingen, Deutschland) wurden im Rahmen eines ADME-panelbasierten Exonsequenzierungsprojektes in den Leberproben umfassend genetische Varianten in den ADME-Genen untersucht. Kodierende sowie flankierende intronische Bereiche und ausgewählte regulatorische Bereiche -insgesamt 4227 Zielsequenzen- aller 340 ADME-Gene wurden simultan mit einem HiSeq 2500 NGS-Gerät (Illumina) von CeGaT sequenziert. Zuvor wurden die spezifischen ADME Regionen über ein „custom enrichment kit“ (Sureselect, Agilent) angereichert und durch einen parallelen Ansatz von beiden Seiten (2×100bp paired-end) mit hoher Sequenziertiefe sequenziert. Die Rohdaten wurden einem Referenzgenom (GRCh37; hg19) zugeordnet, was in 98% der Zielbereiche zu einer mindestens 30-fachen Abdeckung (*coverage*) führte. Die Abdeckung beschreibt die durchschnittliche Anzahl an unabhängigen Sequenzierreaktionen, die einer Base oder einer Region zugeordnet werden konnte.

Für eine CNV-Analyse in NGS-Daten existieren frei verfügbare Programme, die allerdings für genomweite oder exomweite Sequenzierungen optimiert und entwickelt wurden. Deswegen musste ein für eine panelbasierte Sequenzierung angepasstes Verfahren verwendet werden. Im Rahmen dieser Arbeit wurde hierzu die Analyse der Abdeckung (*read depth*) adaptiert (Abbildung 6A). Als Ausgangspunkt dienten BAM-Dateien, welche binär aufgebaut und die Information über die Anlagerung der sequenzierten DNA Abschnitte (*reads*) enthält.

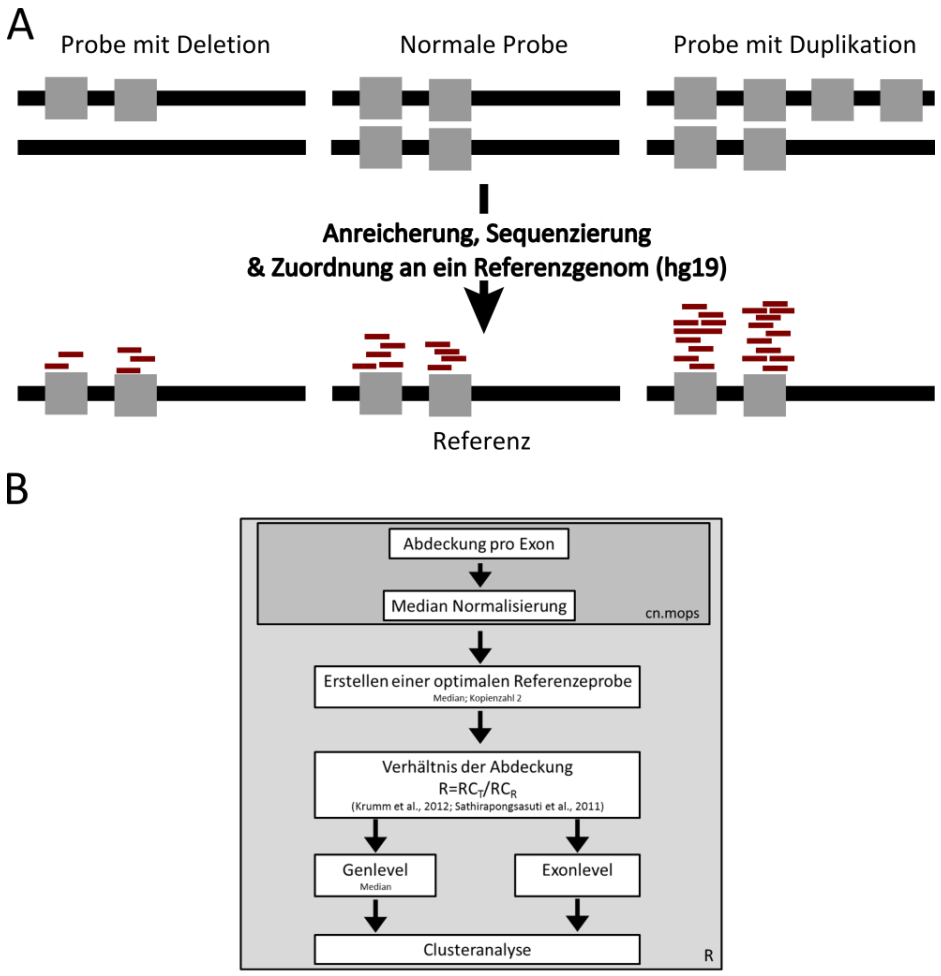


Abbildung 6: Schema zur CNV-Bestimmung in NGS-Daten über die Anzahl der abgelesenen DNA-Abschnitte pro Zielabschnitt. **A** Als graue Rechtecke sind zwei Exons eines Genes dargestellt das in unterschiedlichen Zuständen (deletiert, normal und dupliziert) vorliegt. Nach der Zuordnung der abgelesenen DNA Abschnitte (rote Balken) kann über deren Anzahl Rückschlüsse auf den CNV-Status der Probe gezogen werden. **B** Arbeitsablauf der CNV-Detektion mit der Statistiksoftware R.

Zuerst wurde die Abdeckung, das heißt die absolute Anzahl der reads mit dem cn.mops.1.12.0 Paket in R für jeden sequenzierten Exonbereich ermittelt. Nach einer Mediannormalisierung der Abdeckung wurde für jedes Gen und Probe mit einer zuvor erstellten Referenzprobe ein Verhältnis der Abdeckung berechnet. Die Referenzprobe wurde dabei aus dem Median der zuvor normalisierten Abdeckung erstellt und wurde als Probe mit zwei Kopien definiert. Das Verwenden einer Referenzprobe hat sich in anderen Arbeiten als sehr effizientes Werkzeug der CNV-Bestimmung erwiesen (Krumm et al., 2012; Sathirapongsasuti et al., 2011). Für die CNV-Detektion für jedes Gen (Genebene) wurden alle Exonverhältnisse eines Gens erneut per Median zusammengefasst und in einer Clusteranalyse einer Kopienzahl zugeordnet (V_{Gen}; Abbildung 7A). Um falsch positive Ergebnisse zu vermeiden, wurden zunächst die Werte zentriert und standardisiert (z-Wert) und anschließend ein p-Wert über einen zweiseitigen

Gauß-Test wie folgt berechnet, wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet:

$$p = 2 \cdot \Phi(-|z\text{-}Wert|)$$

Proben mit einem signifikanten Verhältnis (V_{Gen}; p< 0.05) wurden je nach Größe des Verhältnisses in die relativen Kopienzahlgruppen von null (V_{Gen} = 0), eins (V_{Gen} = 0,5), zwei (V_{Gen} = 1), drei (V_{Gen} = 1,5) und größer gleich vier (V_{Gen} ≥ 2) eingeteilt (Abbildung 7B). Lag keine Standardnormalverteilung der z-Werte vor, wurden die Werte manuell einer Kopienzahl zugeordnet. Mit der Annahme, dass ein linearer Zusammenhang zwischen den Verhältnissen (V_{Gen}) und der ermittelten Kopienzahl besteht (Alkan et al., 2009; Sathirapongsasuti et al., 2011) und damit die Steigung einer Regressionsgeraden (s) durch die Gruppen im Idealfall s= 0,5 beträgt, wurden allgemein nur Gene mit CNV-Treffer aufgenommen, deren Regressionsgerade eine Steigung von mehr als s≥ 0,4 betrug (Abbildung 7C). Abschließend wurde die Clusteranalyse für jedes Exon eines Gens wiederholt und dabei die gleichen Filterkriterien und Einstellungen angewendet. Eine Übersicht des Arbeitsablaufs findet sich in Abbildung 6B.

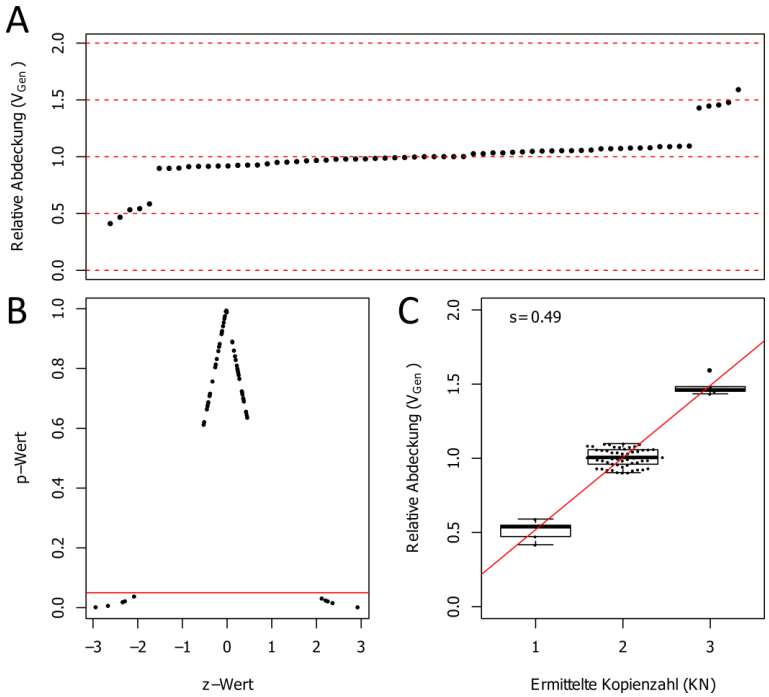


Abbildung 7: Beispieldatensatz (n= 60) zur Verdeutlichung der angewendeten CNV-Analysemethode. **A** Clusteranalyse der relativen Abdeckung. Das Verhältnis berechnet sich aus der Anzahl an sequenzierten DNA Abschnitten (engl. *Reads*) von Probe und der zuvor gebildeten Referenzprobe. **B** Aufgetragen ist der berechnete p-Wert gegen den zentrierten und standardisierten z-Wert. Für Proben mit signifikantem z-Wert (Die horizontale rote Linie verdeutlicht das Signifikanzniveau p= 0,05) wurde die entsprechenden Kopienzahl ermittelt. **C** Mit der ermittelten Kopienzahl und dem Verhältnis der Abdeckung wird eine Regressionsgerade und deren Steigung (s) berechnet. Die Gerade ist als rote Linie im Schaubild eingezeichnet.

2.2.4.2 SNP-Genotypen des ADME-Exonsequenzierungspanels

Das Auffinden und die Annotation (dbSNP build 137) von SNPs und Indels (kleine Insertionen und Deletionen) wurde von CeGaT Mitarbeitern mittels ihrer bioinformatischen Pipeline vorgenommen und die Daten am IKP in Genedata (Version 9) gespeichert. Mit der Analyst Software wurden detektierte SNP-Genotypen ausgeschlossen, wenn an der SNP-Position die Abdeckung kleiner 20 oder die Variationsfrequenz (VAF) unter 20% lag. Diese Grenzwerte wurden von CeGaT empfohlen und für diese Arbeit übernommen. Dabei beschreibt die VAF den prozentualen Anteil an sequenzierten DNA Segmenten, die eine bestimmte Variation tragen. Des Weiteren wurden SNPs ausgeschlossen, die in allen Proben homozygot vorlagen. Außerdem musste in mindestens 75 Proben ein SNP-Genotyp vorliegen und sich im Hardy Weinberg Equilibrium (HWE) befinden ($p \geq 0{,}001$). Das HWE wurde mit der summary-Funktion des SNPassoc_1.9-2 Paket in R exakt berechnet. Zuletzt wurden nur SNPs, deren Positionen innerhalb der von CeGat amplifizierten Zielbereiche lagen, berücksichtigt. Nach dem Filtern standen Daten zu 6.884 SNPs für die Analyse zur Verfügung.

2.2.4.3 TaqMan, eine quantitative PCR-Methode

Eine quantitative Polymerasenkettenreaktion (qPCR) zeichnet sich neben der herkömmlichen Vervielfältigungsmethode für Nukleinsäuren (Mullis and Faloona, 1987) durch eine zusätzliche Quantifizierung der amplifizierten cDNA aus. Die TaqMan-Methode beruht auf der Hybridisierung einer TaqMan-Sonde (auch Hydrolyse-Sonden) innerhalb des durch spezifische PCR-Primer amplifizierten DNA Abschnittes. Die Sonde trägt am 5‘-Ende einen Fluoreszenzfarbstoff sowie am 3‘-Ende einen Quencher, der die Energieabgabe des Farbstoffes als Strahlung durch einen physikalischen Prozess (Förster-Resonanzenergietransfer) verhindert. Während der Elongation werden angelagerte Sonden durch die 5‘-Exonuklease-Aktivität der Taq-DNA-Polymerase hydrolysiert und somit der Quencher vom Flurophor räumlich getrennt. Die Intensität der Fluoresenz steigt somit an und ein Signal detektiert. Durch die Verwendung von unterschiedlichen Farbstoffen können mehrere Assays in derselben Probe simultan gemessen werden. Diese Methode wurde benutzt, um die DNA Kopienzahl und die mRNA Expression quantitativ zu messen und ist am IKP etabliert (Schaeffeler et al., 2003).

Bestimmung der Kopienzahl

Speziell für die CNV-Analyse entwickelte und vorgefertigte 20x TaqMan CNV-Assays (Applied Biosystems) wurden dazu genutzt, ausgewählte CNVs in ADME-Genen zu quantifizieren (siehe Material). Vorgegangen wurde nach Herstellerangaben. Die Assays für spezifische ADME-Genbereiche enthalten neben den Primern Sonden, die mit

dem Farbstoff 6-carboxyfluorescein (FAM) markiert waren. Als Referenzassay wurde eine Sonde, die in einem DNA Abschnitt des *RNase P* Gens bindet und mit dem Farbstoff VIC (keine Angaben zur chemischen Struktur bekannt) markiert ist, verwendet. Die *RNase P* Genregion wird als CNV-freier Locus und in allen Proben mit der Kopienzahl zwei vorkommend angenommen. Beide Assays und ~10ng genomische DNA wurden in 96er Mikrotiterplatten in einem 7500 qPCR System (Applied Biosystems) in biplex Reaktionen verwendet. Ein Vorteil der gleichzeitigen Amplifikation von Probe- und Referenzbereich in derselben Probe ist die Minimierung von Pipettierfehlern und Unterschieden der Effizienz, mit der die Amplifikation der DNA Bereiche voranschreitet.

Tabelle 5: MasterMix für qPCR-Genotypisierungen

Reagenz	**Menge**
qPCR Master Mix	6
TaqMan Assay	0,6
RNase P TaqMan Assay	0,6
H_2O	4,8

Das Gerät wurde mit Standardeinstellungen betrieben (C_T Grenzwert 0,2 mit automatischer Basislinie; Der C_T-Wert entspricht der Anzahl an PCR-Zyklen bis ein Fluoreszenzanstieg über einen definierten Schwellenwert erreicht ist). Das Experiment wurde in Duplikaten und gegebenenfalls in Triplikaten vorgenommen. Die Auswertung erfolgte mit der Software Copy Caller v.2.0 Software (Applied Biosystems) und basierte auf der $\Delta\Delta C_T$-Methode.

Quantifizierung der mRNA Expression

Aus Lebergewebe wurden alle RNA Moleküle mit dem Trizol® (Invitrogen, CA, USA)/ Qiagen RNeasy® Protokol wie beschrieben isoliert (Gomes et al., 2009). Die Expression von ausgewählten ADME-Genen wurde mit TaqMan Expressions-Assays und dem TaqMan 7500 Gerät (Applied Biosystems) bestimmt. Die Rohdaten wurden gegen *RPLP0* (60S large ribosomal protein P0), dessen Expression mit dem endogenen Assay 4326314E (Applied Biosystems) gemessen wurde, normalisiert.

2.2.4.4 ADME-weite Genexpression

In einer vorausgegangenen Arbeit wurde in allen Leberproben die genomweite mRNA Expression mit dem Mikrochip Illumina Human WG6.V2 gemessen (Schröder et al., 2013). Aus diesen Daten wurden die Expressionslevel für alle 340 ADME-Gene extra-

hiert. Wenn mehr als eine Sonde für ein Gen verfügbar war, wurden die Sondensequenzen mit BLAST (NCBI, Bethesda MD, USA) im humanem Genom (hg38.p2) relokalisiert und die Sonde mit der optimalsten Lage ausgewählt.

2.2.4.5 WesternBlot Analyse zur Proteinbestimmung in humanen Lebern

Im Rahmen dieser Arbeit wurde die SULT1A1 Proteinmenge in den IKP148-Leberproben bestimmt. Obwohl die CYP2E1 Proteindaten bereits vorlagen, wird die Methode der Vollständigkeit halber in dieser Arbeit dokumentiert.

Die Proteinfraktion aus Zytosol (SULT1A1) oder Lebermikrosomen (CYP2E1) wurde durch eine SDS-Polyacrylamidgelelektrophorese aufgetrennt. Anschließend wurde die Proteinmenge mit der WesternBlot Methode und einem immunochemischen Nachweis bestimmt.

Die Proben wurden mit Lämmli Puffer versetzt und für 3-5 min bei 95°C denaturiert. Durch die Behandlung werden Sekundär- und Tertiärstrukturen aufgebrochen und Disulfidbrücken durch das im Puffer enthaltene β-Mercaptoethanol reduziert und gespalten. Das anionische Detergenz SDS (Natriumdodecylsulfat) belädt die linearisierten Proteine im Verhältnis zu ihrer Masse mit negativer Ladung. Dadurch ist eine Proteinauftrennung nach Masse im elektrischen Feld möglich.

25µg denaturiertes Proteingemisch wurde auf ein 10% (CYP2E1) oder 11% (SULT1A1) SDS-Polyacrylamid Gel aufgetragen. Nach erfolgreicher Auftrennung (150V, 4-5h) wurde das Gel elektrophoretisch auf eine Nitrozellulosemembran in einer Fastblot Semi-Dry Apparatur (Biometra) 15 Minuten und bei 400mA transferiert. Eine Ponceau-S-Färbung diente zur Überprüfung der Probenbeladung sowie der Effizienz des Transfers. Die über Nacht getrocknete Membran wurde am nächsten Tag für den immunochemischen Proteinnachweis in 5% Magermilchpulver in TBST-Puffer 1h bei Raumtemperatur (RT) geblockt. Die Inkubation der Membran mit dem Primärantikörper erfolgte in einer 1% (CYP2E1) oder 5% (SULT1A1) Magermilch/TBST Lösung für 40min (SULT1A1) bis 1h (CYP2E1) bei RT. Nach drei 10 minütigen Waschschritten in TBST wurde die Membran 30min mit dem sekundären Antikörper in einer Verdünnung von 1:10 000 in 1% bzw. 5% Magermilch/TBST-Puffer behandelt. Nach drei weiteren 10 minütigen Waschschritten konnten die spezifischen Proteinbanden detektiert werden. Für CYP2E1 wurde ein chemilumineszenter Nachweis über einen sekundären Antiköper, der mit einer Peroxidase gekoppelt war, einem Signal Substrat und der Fuji LAS-1000 CCD Kamera (Raytest, Straubenhardt, Germany) verwendet. Die Detektion der SULT1A1 Proteinbande erfolgte durch einen mit einem Fluoreszenzfarbstoff gekoppelten Antikörper und dem Odyssey® Infrared Imaging System (LI-COR

Biosciences). Die absolute oder relative Proteinmenge wurde über eine auf jeder Membran mitgeführten Verdünnungsreihe von rekombinant exprimiertem Protein (CYP2E1) oder gepoolten Proteinproben (SULT1A1; 5µg, 10µg, 20µg und 40µg) berechnet. Alle Analysen wurden mindestens zweimal wiederholt. Die relativen SULT1A1 Werte wurden auf die Probe mit niedrigster Expression normalisiert.

2.2.4.6 Analytische Methoden

Quantifizierung der CYP Enzymaktivität

Die Aktivitätsdaten von ausgewählten CYPs waren bereits vorhanden. Die Methode wurde erstmalig in Gomes et al. (2009) beschrieben. Aus Lebergewebe isolierte Mikrosomen wurden mit spezifischen Substraten einzeln oder simultan mit dem sogenannten CYP-Cocktail Assay inkubiert und anschließend wurden die Produkte im Medium per Massenspektroskopie ausgewertet und quantifiziert. Substrat für CYP2A6 war Coumarin (200µM), für CYP2D6 Propafenon (5µM) und für CYP2E1 Chlorzoxazon (75µM).

Messung von Methyleugenol DNA-Addukten

DNA-Addukte wurden mit dem publizierten Verfahren (Herrmann et al., 2013) der Flüssigchromatographie mit einem gekoppelten Massenspektrometer am Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke (DIfE) vom Kooperationspartner in 121 der 150 humanen Leberproben gemessen.

2.2.4.7 Imputierung von unbeobachteten SNP-Genotypen und genomweite Assoziationanalyse

Das Verfahren der Imputierung von SNP-Genotypen ermöglicht eine umfassende genomweite Assoziationsanalyse. Ausgehend von genomweiten SNP-Daten können durch einen Vergleich mit Haplotypen einer Referenzkohorte nicht beobachtete SNPs in der Studienkohorte mit hoher Wahrscheinlichkeit vorhergesagt werden (Abbildung 8). Mehr als 300.000 SNP-Genotypen waren für jede Leberprobe verfügbar (Schröder et al., 2013). Vor der Imputierung wurden SNPs mit schlechter Qualität mit der Analysesoftware PLINK entfernt. Die Allelefrequenz (MAF< 1%), fehlende Genotypen eines SNPs (Detektionsrate < 90%) und HWE (p-Wert< 10^{-6}) mussten erfüllt sein. Ferner wurde die Position und die rs-Nummer des SNPs auf das Referenzgenom GRCh37 (hg19) und der Version dbSNP141 aktualisiert. Dazu wurden die NCBI Archivdateien SNPHistory.bcp, SNPHistory.bcp und b141_SNPChrPosOnRef_GRCh37p13.bcp (http://goo.gl/BTg0WP) und das R Zusatzpaket biomaRt_2.20.0 (Durinck et al., 2009) benutzt.

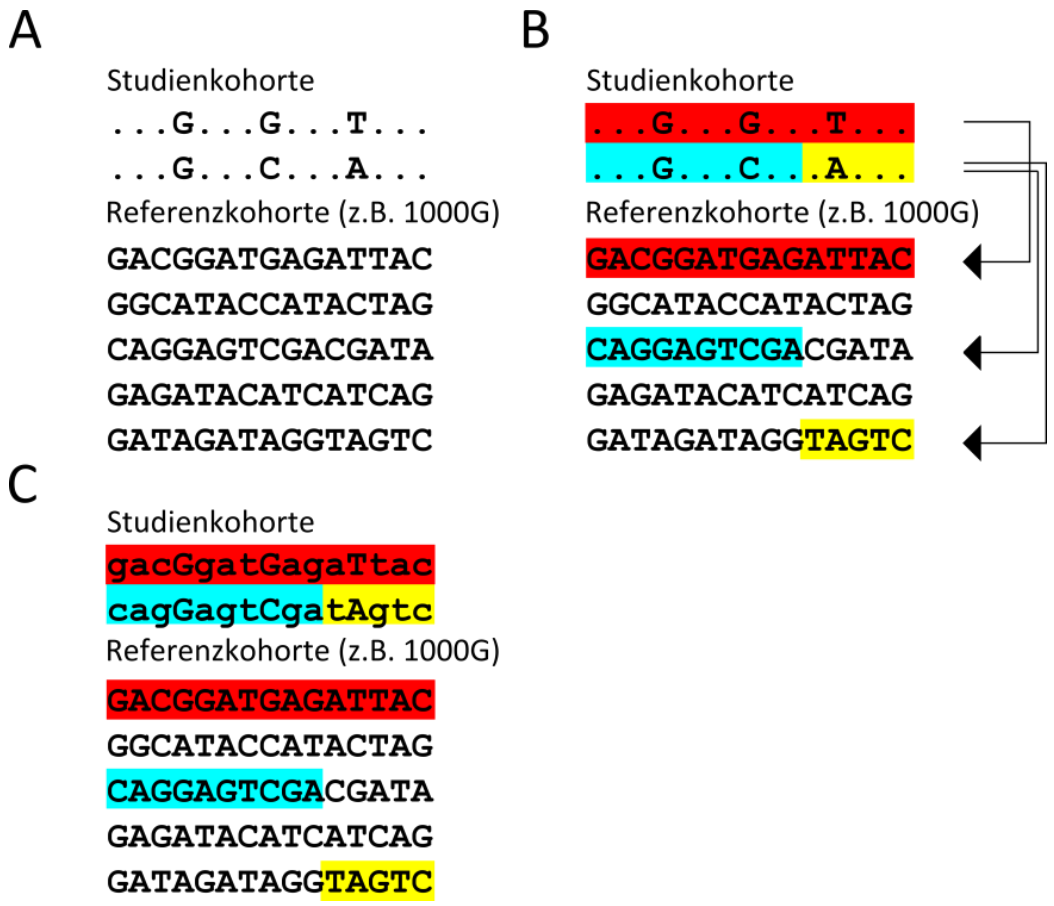


Abbildung 8: Ablauf der Imputierung **A** Studie mit zwei Proben und einer kleinen Anzahl an bekannten SNP-Genotypen und einer Referenzkohorte aus Proben mit umfassender Genotyp- und Haplotyp-Information. **B** Suche nach gleichen SNP-Mustern in der Studien- und Referenzkohorte. **C** Auffüllen der fehlenden Genotypen (kleine Buchstaben) in der Studienkohorte. Abbildung adaptiert aus Li et al., (2009).

Insgesamt konnten für 149 Proben 305.111 autosomale SNPs mit hoher Qualität der Imputierungspipeline übergeben werden. Zuerst wurden mit ShapeIT (Delaneau et al., 2012) die Haplotypen bestimmt. Mit der Software Impute2 v2.3.1 (Howie et al., 2012) wurden diese mit den Haplotypen des Referenzpanel verglichen und fehlende SNPs eingefügt. Als Referenz dienten Daten des 1000Genom Projektes (1,000 Genomes haplotypes, Phase I integrated variant set; release June 2014; hg19). Nach erfolgter Imputierung wurden Varianten mit niedriger Imputierungsgüte entfernt (Wahrscheinlichkeitswert < 0.8 und Infowert <0.8.). Die im Abschnitt 2.2.4.1 bestimmten CNVs wurden anschließend hinzugefügt. Dabei wurde die Kopienzahl in eine SNP-Kodierung überführt. Eine Genkopie wurde als AA= +/- festgelegt. Zwei Kopien wurden als A/T= +/+ und gleich oder mehr als drei Genkopien wurden als T/T= ++/+ definiert. Anschließend wurde mit SNPTEST v2.5 (Marchini et al., 2007) und der „frequentist“-Test Option (Expected-Methode) eine Assoziationsstudie mit den SULT1A1 Proteindaten als Endpunkte durchgeführt. Da dieser Test einer linearen Regression ähnlich ist, wurde die Verteilung der SULT1A1 Proteinwerte mit einem Histogramm, einem Quantil-

Quantil-Diagramm, sowie dem Shapiro-Wilk Test erfolgreich auf Normalverteilung überprüft. Ein Manhattan-Diagramm der erhaltenen p-Werte wurden mit Hilfe des qqman_0.1.2 Paketes (Turner, 2014) erstellt. Das genomweite Signifikanzniveau wurde bei p= 5E-08 festgelegt. Dies entspricht einer Bonferroni-Korrektur bei einer Million unabhängigen Tests. Ein Suggestivniveau, welches ein falsch Positives Ergebnis in der Analyse zulässt, wurde bei 1E-06 (1/1E06) gewählt.

2.2.4.8 Analyse des Kopplungsungleichgewichts zwischen *CYP2E1* Duplikationen und SNPs

Die Analyse von SNPs, die über-zufällig häufig gemeinsam mit der *CYP2E1* Duplikationen in einer Population vorkommen, wurde wie folgt berechnet: Der CNV-Status der zusätzlichen Genkopien wurde als SNP-Genotyp kodiert, wobei zwei Kopien als ein AA, drei Kopien als AT und vier Kopien als TT definiert waren. SNP-Genotypen in der *CYP2E1* Region auf dem Chromosomenarm 10q26.3 (hg19: 135 200kb-135 500kb) wurden über das 1000 Genomprojekt bezogen (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/). Das gepaarte LD (r^2) zwischen *CYP2E1* Duplikation und SNPs wurde mit der Software PLINK berechnet.

2.2.4.9 Weitere statistische Analysen

Der Spearmans Rangkorrelationskoeffizient wurde dazu benutzt, den Zusammenhang zwischen zwei Variablen (mRNA, Protein und Aktivitätsdaten sowie dem nicht-genetischen Faktor Alter) zu beschreiben. Für eine univariate Genotyp-Phänotyp Analyse wurde der gepaarte Wilcoxon-Vorzeichen-Rang-Test oder der Kruskal-Wallis-Test verwendet. Eine multivariate Analyse wurde wie oben bereits geschrieben mit einem linearen (gegebenenfalls mit transformierten Endpunkten) oder mit einem medianen Regressionsmodell in R durchgeführt. In SNP-Analysen in denen drei verschiedene genetische Modelle (dominant, rezessiv und additiv) der Variante untersucht wurden, gruppierte das SNPassoc_1.9-2 Paket die Proben entsprechend deren Genotyp. Im additiven Modell wurden die Proben einer numerischen Variable (1, 2, 3) zugeordnet. Im dominanten und rezessiven Modell wurden die Proben jeweils in zwei Gruppen einsortiert (dominant, AA vs. Aa & aa; rezessiv, AA & Aa vs. aa; „A“ ist als das Hauptallel und „a“ als das alternative Allel definiert). Alle Tests waren zweiseitig und das Signifikanzniveau lag bei $p<0,05$. Das Benjamini-Hochberg Verfahren wurden für die Korrektur der p-Werte bei multiplen Testen verwendet (Benjamini and Hochberg, 1995).

3 Resultate

3.1 Systematische Suche nach Kopienzahlvariationen von ADME-Genen

3.1.1 Auswertung einer öffentlichen Datenbank für genomische strukturelle Varianten

Auf die Daten in DGV wurde online über einen Genombrowser und per FTP-Client zugegriffen. Wie bereits in der Einleitung erwähnt, wurden nur Gene als CNV-assoziiert angenommen, wenn mehr als zwei CNVs von mindestens zwei unterschiedlichen Studien in der Genregion detektiert wurden. Dieses Vorgehen sollte falsch positive Ergebnisse minimieren.

3.1.1.1 Allgemeine CNV-Verteilung im humanen Genom

Zum Zeitpunkt dieser Untersuchung (Zugriff am 16.10.2014) waren in DGV insgesamt 353.126 CNVs aus 62 Studien katalogisiert. Die chromosomale Verteilung war dabei unterschiedlich. So wurden nur in 29,4% der Y-chromosomalen Sequenzen CNVs detektiert, wohingegen in 88,2% der Sequenzen auf Chromosom 7 CNVs beschrieben wurden. Insgesamt wurden in durchschnittlich 67% des Genoms CNVs gefunden (Abbildung 9A). Die meisten der beschriebenen CNVs wiesen eine Länge von 1-10kb auf (Abbildung 9B). Betrachtet man nur kodierende Regionen ergab sich folgendes Bild: Insgesamt wurden in ca. 91% der ~41.516 Transkriptregionen und 80% aller miRNA-Loci CNVs detektiert.

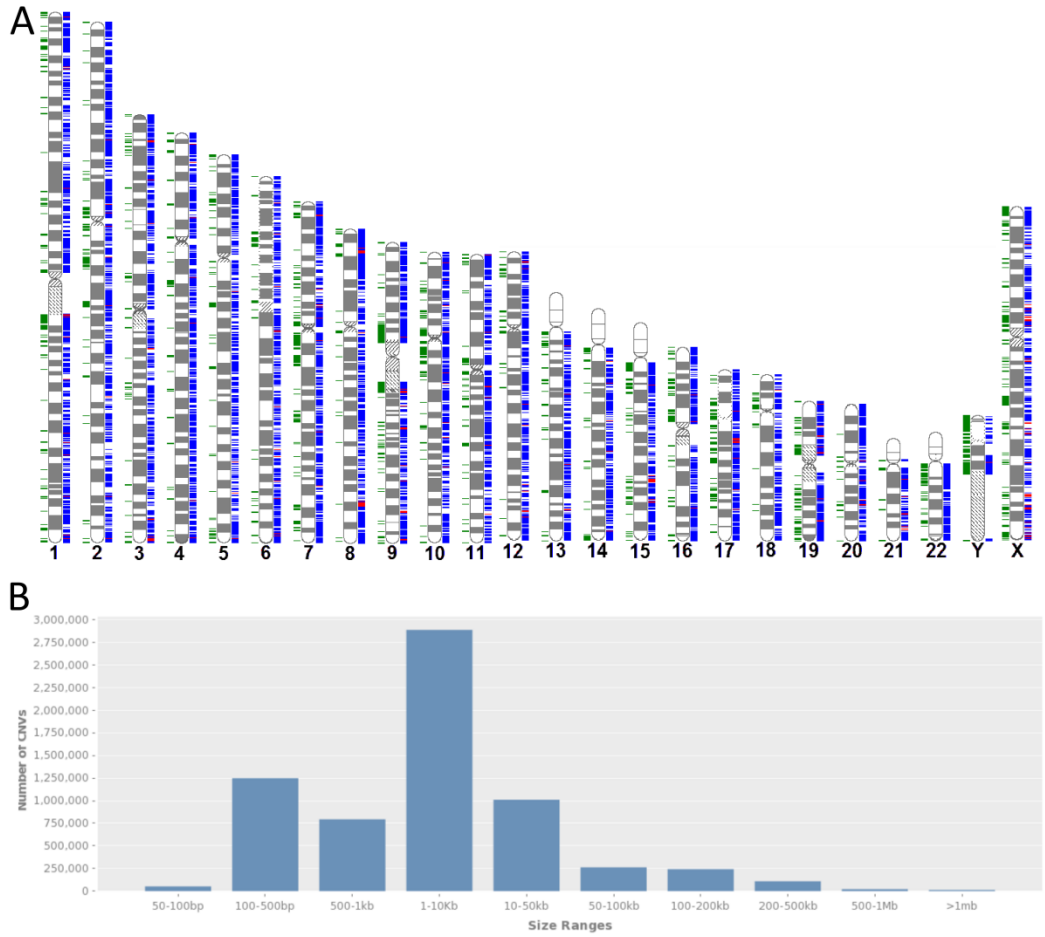


Abbildung 9: **A** Genomweite Übersicht der in DGV eingetragenen CNVs. Blaue Balken zeigen CNVs, rote Banden Inversionsbruchpunkte und in Grün sind segmentale Duplikationen dargestellt. **B** Histogramm der Länge aller in DGV gelisteten CNVs. Übernommen aus DGV (MacDonald et al., 2014).

3.1.1.2 Allgemeines CNV-Vorkommen in ADME-Genen in DGV

Die Genloci der 340 ADME-Gene wurden mit den Start- und Endpunkten der in der Datenbank registrierten CNVs abgeglichen. Nur CNVs, die die Filterkriterien erfüllten (siehe 2.2.1) und das Zielgen komplett einschlossen, wurden in der Analyse berücksichtigt. In den ADME Regionen wurden insgesamt 13.912 Deletionen und 2.793 Duplikationen gefunden, die sich auf 25% der Phase I, 33% der Phase II, 30% der Transporter, 12% der Modifizierer und 32% der ADME-verwandten Genregionen aufteilten (Tabelle S1). Darunter befanden sich neben den pharmakologisch relevanten Genen mit bereits beschriebenen CNVs wie z.B. *CYP2D6*, *CYP2A6*, *GSTT1*, *GSTM1*, *UGT2B28* und *UGT2B17*, auch weitgehend unbekannte CNVs in Genen der Phase I und II wie *CYP2E1*, *UGT2B15*, *UGT2B11* und verschiedenen Transportern (Tabelle S1). Bemerkenswert war der Toptreffer *CYP2E1*, in dessen Genregion von 13 Studien 40 CNVs gefunden wurden. Allerdings traten diese selten auf (Deletion= 1%; Duplikation= 5%). Die Frequenzen zwischen Gewinn- und Verlustvarianten unterschieden sich in den ADME-Gruppen. So waren Duplikationen in der Gruppe der Phase I und II Gene

häufiger (durchschnittliche Frequenz von 3% und 6%) als Duplikationen in Transportern, ADME-Verwandten und Modifizierer (durchschnittliche Frequenz jeweils kleiner als 1%). Deletionen waren weit häufiger und fanden sich mit einer jeweiligen durchschnittlichen CNV-Frequenz von 12% in Phase I, 15% in Phase II, 7% in Transportern und ADME-verwandten Genen und 14% in der Gengruppe der Modifizierer. Wie in Abbildung 10 zu erkennen ist, waren die meisten CNVs selten (<5%). Deletionen mit Frequenzen größer als 30% wurden in den Genen *GSTT1*, *GSTM1* und *UGT2B28* gefunden (Tabelle S1).

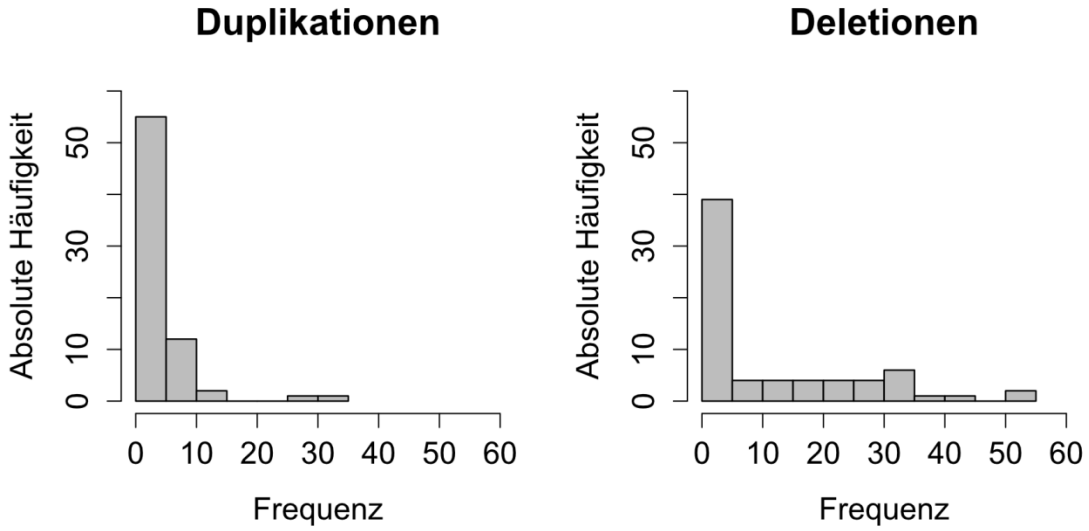


Abbildung 10: Absolute Häufigkeiten der CNV-Frequenzen.

3.1.1.3 Populationsabhängige ADME-weite CNV-Verteilung

Aus Populationsstudien von SNP-Datensätzen war bekannt, dass sich Allelfrequenzen zwischen unterschiedlichen Populationen stark unterscheiden können. Deswegen wurde die Analyse auf Proben des HapMap-Projektes, die sich einer bestimmten Population zuordnen ließen, beschränkt. In 596 Proben mit afrikanischer, asiatischer und europäischer Abstammung wurden CNVs in 19% der Phase I, 24% der Phase II Gene, 11% der Transporter, 4% der ADME-Verwandten und 3% der Modifizierer detektiert. In Proben mit afrikanischer Herkunft (YRI, n=195) wurden durchschnittlich 4,64 CNVs pro Probe gefunden. In Proben aus dem asiatischem Raum (CHB & JPT; n=217 Proben) waren es 4,34 CNVs pro Donor. Proben mit europäischer Abstammung (CEU, n=184) trugen durchschnittlich 4,26 CNVs in ADME-Genen. Die marginalen Unterschiede zwischen der europäischen Population und den anderen waren signifikant ($p< 0.006$). Wie in Tabelle 6 gezeigt, war das Vorkommen der CNVs in den einzelnen ADME-Gruppen je Population unterschiedlich. Zum Beispiel wiesen 11% und 13% der Phase I Gene CNVs in YRI- und CEU- Proben auf. Im Gegensatz dazu fanden sich in 5% in der asia-

tischen Population CNVs der Phase I Gene. *CYP2D6* und das *FMO* Gencluster waren im Vergleich überhaupt nicht betroffen. Außerdem wurden seltene CNVs in Genen der Gruppe Modifizierer nur in Proben aus Asien und Afrika gefunden. Insgesamt waren mehr Phase I und II Gene von CNVs betroffen als Transporter oder Modifizierer wie nukläre Rezeptoren. Gleichzeitig waren diese CNVs häufiger (f>1%). CNVs in Transportern waren überwiegend selten (f≤ 1%).

Tabelle 6: Relativer Anteil an CNV-assoziierten Genen in der jeweiligen ADME-Gengruppe in HapMap-Proben (n=596). Benutzt wurden CNV-Informationen aus 10 Studien der DGV-Datenbank (siehe Tabelle S2).

Frequenz	**Population[1]**	**Phase I (n=96)**	**Phase II (n=51)**	**Transporter (n=110)**	**Modifizierer (n=55)**	**ADME-Verwandte (n=28)**
≤ 1%	CEU	3%	4%	4%	-	-
	CHB &JPT	1%	4%	4%	4%	4%
	YRI	2%	2%	4%	2%	-
> 1%	CEU	9%	18%	-	-	-
	CHB &JPT	4%	14%	1%	-	-
	YRI	9%	22%	-	-	-
Insgesamt	CEU	13%	22%	4%	-	-
	CHB &JPT	5%	18%	5%	4%	4%
	YRI	11%	24%	4%	2%	-

[1]HapMap-Proben aus unterschiedlichen geografischen Regionen. YRI= Yoruba aus Ibadan, Nigeria (n= 195); CHB= Han Chinese aus Beijing (n= 112), China; JPT= Japaner aus Tokyo, Japan (n= 105); CEU= Einwohner nord- und westeuropäischer Vorfahren aus Utah, USA (n= 184).

Die durchschnittliche CNV-Häufigkeit lag bei 11%. In Proben mit asiatischer Abstammung war die CNV-Frequenz durchschnittlich höher (14%) als in den anderen zwei Populationen (CEU: 11%; YRI: 10%). Träger einer Deletion waren generell häufiger (YRI f= 13%; CEU f= 19%; CHB & JPT f= 24%) als Träger einer Duplikationen (YRI f= 4%; CEU f= 2%; CHB & JPT f= 7%). Wie in Abbildung 11 zu sehen ist, ließen sich hochfrequente Deletionen in der Gruppe der Phase II Gene bestätigen. Insbesondere Gene der GST- und UGT- Familie, wie zum Beispiel *GSTT1*, *GSTM1* und *UGT2B17* und *UGT2B28* zeigten Deletionshäufigkeiten von jeweils über 30-50%. Auffällig waren in diesen Genen auch bisher unbeschriebene Duplikationen, die aber nach näherer Betrachtung als falsch positive Resultate eingeordnet werden mussten. In den betroffenen Studien wurde ein relatives CNV-Analyseverfahren verwendet (Oligo-

Mikrochips), welches aus dem Vergleich einer Probe und einer Referenz (NA10851 oder NA15510) die Kopienzahl ermittelte. Aus dem Vergleich aller Quellen konnte eine Deletion in den Genen *GSTM1*, *GSTT1*, *UGT2B28* oder *UGT2B17* in ebendiesen Referenzproben nachvollzogen werden. Deswegen wurde beim Vergleich einer Probe mit zwei Genkopien gegen die Referenzprobe mit einer Kopie fälschlicherweise drei Genkopien vorhergesagt (Abbildung 11, gekennzeichnet mit einer Raute). Auffällig war, dass keine strukturellen Varianten in Proben aus dem asiatischen Raum für *CYP2D6* oder *SULT1A1* gefunden wurden und auch in den anderen zwei Populationen das Vorkommen von Deletionen und Duplikationen in beiden Genen im Vergleich zu bekannten Daten selten war.

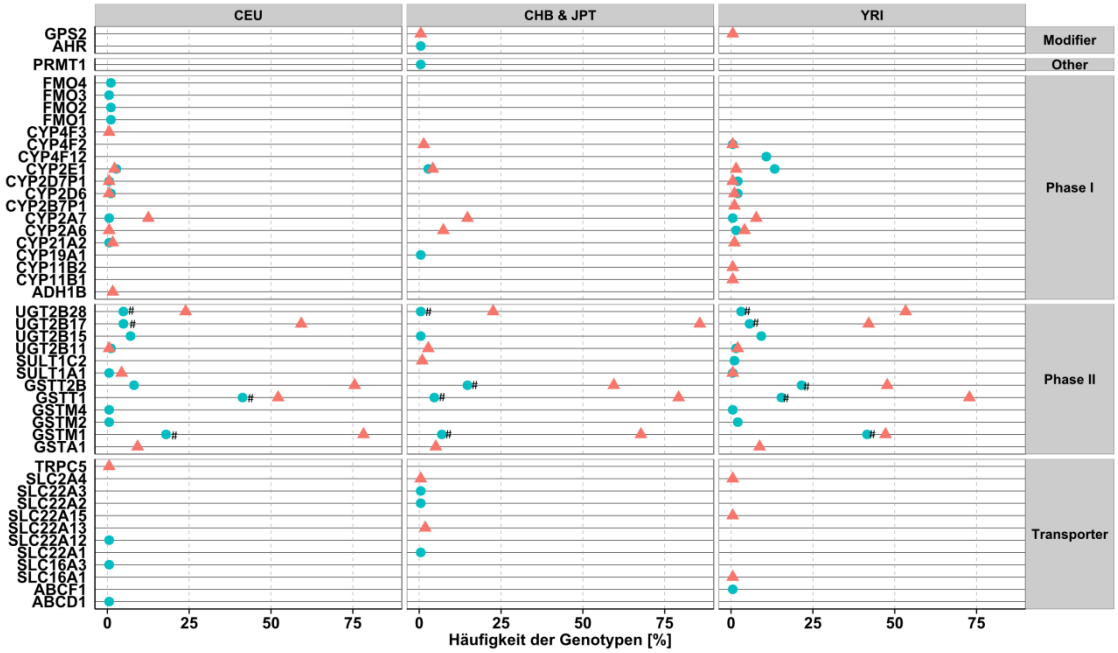


Abbildung 11: Übersicht über die Häufigkeiten der ADME-CNV Genotypen in DGV im Populationsvergleich. Blaue Kreise zeigen Duplikationen, rote Dreiecke Deletionen. Datengrundlage sind CNV-Informationen in DGV (Stand Oktober 2014) aus HapMap-Proben (CHB= Han Chinese aus Beijing, China; n= 112; JPT= Japaner aus Tokyo, Japan; n= 105; YRI= Yoruba aus Ibadan, Nigeria; n= 195; CEU= Einwohner mit Nord- und Westeuropäischer Vorfahren aus Utah, USA; n= 184) von 10 Studien (siehe Tabelle S2). Mit einer Raute sind falsch positive strukturelle Varianten gekennzeichnet. Das untersuchte Gen wies in der Referenzprobe keine Kopienzahl von zwei aus.

3.1.2 CNV-Auswertung in Leberproben des TCGA-Projektes

Die Ergebnisse in dieser Arbeit basieren auf Daten, die vom TCGA-Forschungsnetzwerk produziert wurden (http://cancergenome.nih.gov/). In der Datenbank des TCGA-Projektes findet man Phänotypisierungs- und Genotypisierungsdaten von gepaarten Biopsien aus hepatozellulären Karzinomen (HCC) sowie aus gesundem

Lebergewebe. Für das Leberkarzinom Projekt sind genomweite CNV- und Genexpressionsdaten, sowie klinische und demografische Informationen zu den jeweiligen Geweben und Donoren frei zugänglich und analysierbar.

3.1.2.1 Deskriptive Analyse der genomweiten CNV-Verteilung in humanen TCGA-Lebergewebe

Im gesunden Gewebe der Leberdonoren (n= 348) des TCGA-Konsortiums konnten insgesamt 36.767 Segmente (autosomale Varianten) identifiziert werden die als CNV definiert wurden und von denen 94% mit strukturellen Varianten, die bereits in DGV beschrieben waren, überlappten (siehe Methode 2.2.3.1). Der Vergleich der bestimmten CNVs in dieser Arbeit mit CNV-Treffern einer Studie in DGV, die den gleichen SNP-Mikrochip verwendete (Altshuler et al., 2010) zeigte, dass die berechneten Längen der CNVs signifikant übereinstimmten (Abbildung 12A; rs=0.92; p< 2.2e-16). Die mediane Abweichung der Start und Endpunkte betrug dabei 2kb, wobei die meisten CNVs nur eine 1kbp Abweichung von Start- und Endpunkt aufwiesen (Abbildung 12B).

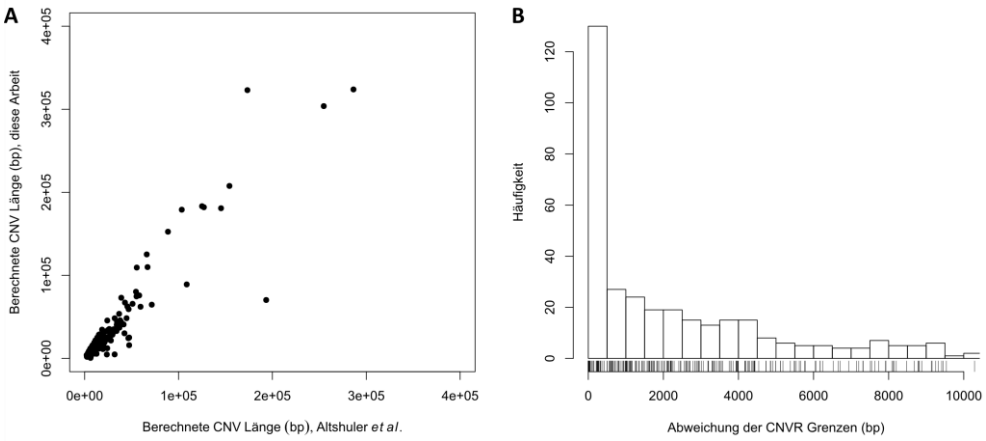


Abbildung 12: **A** Vergleich der ermittelten Längen der CNV-Regionen dieser Arbeit und Altshuler et al. (2010) **B** Verteilung der Abweichung des Betrags von Start und Endpunkten der vergleichbaren CNVs.

Pro Patient wurden durchschnittlich 106 CNVs gefunden, wobei signifikant (Wilcoxon-Vorzeichen-Rang-Test; p< 2e-16) mehr Deletionen ($n_{Mittelwert}$= 62) als Duplikationen ($n_{Mittelwert}$ = 44) beobachtet wurden. Im Tumorgewebe war die CNV-Anzahl signifikant (gepaarter t.test p< 2.2e-16) erhöht. Im Gegensatz zum gesunden Gewebe unterschied sich interessanterweise im Tumor die Anzahl der Deletionen (n= 106) und der Duplikationen (n= 109) pro Spender nur noch minimal (Abbildung 13B). Im Tumor war eine generelle Zunahme von langen strukturellen Varianten (Länge> 1000kb) zu beobachten (Abbildung 13A & C). Am signifikantesten war die Zunahme von langen Dup-

likationen und in den Chromosomen 1,5,6,8 und 20 (> 2,2-fache Zunahme; Abbildung 13A & C).

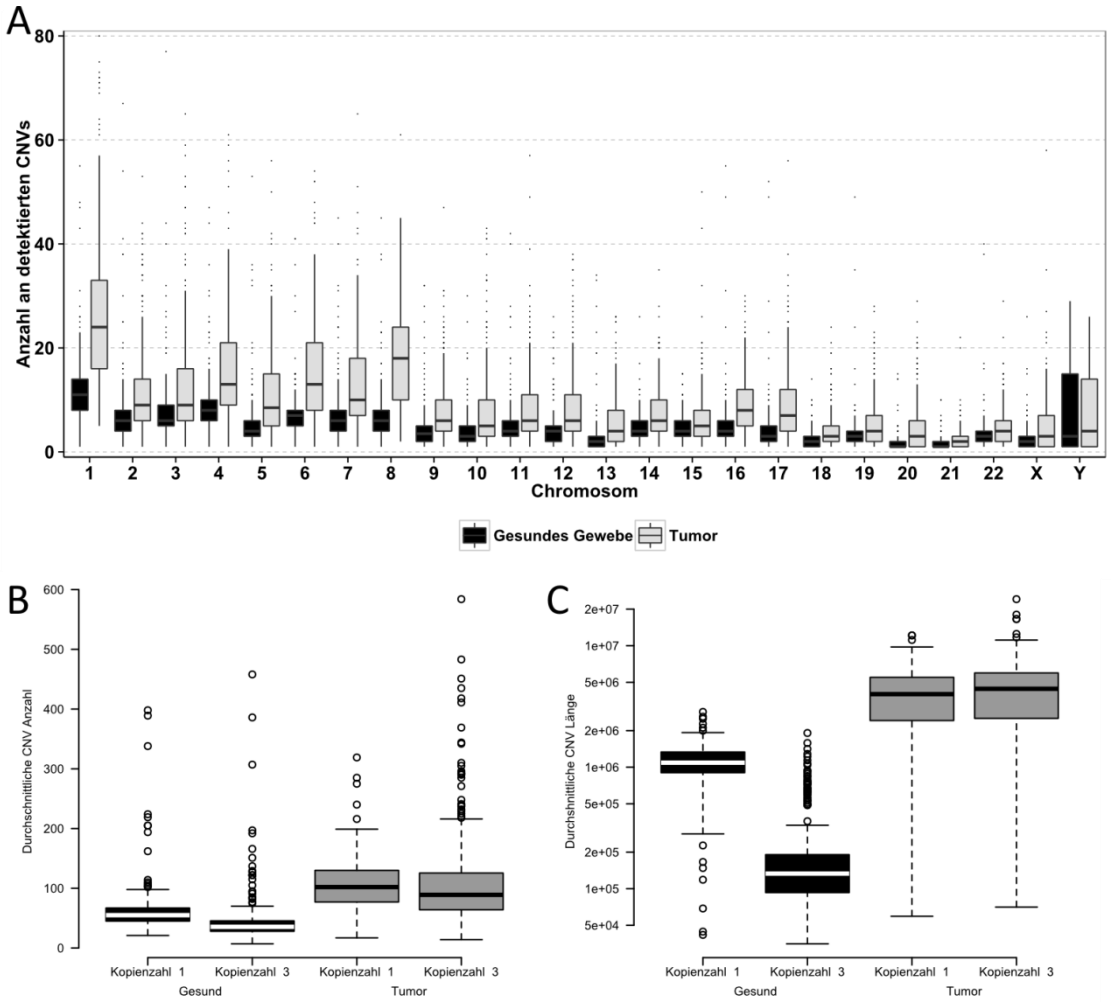


Abbildung 13: CNV-Verteilung und Länge in den TCGA-Leberproben. **A** Vorkommen der CNVs pro Chromosom im gepaarten Vergleich von gesundem Lebergewebe und hepatozellulärem Karzinomgewebe. **B** Durchschnittliche CNV-Anzahl und **C** CNV-Länge je TCGA-Probe. Die Proben wurden nach CNV-Art und Gewebetyp gruppiert.

3.1.2.2 CNVs von ADME-Generegionen im TCGA-Datensatz

Ein Abgleich der Positionen der 340 untersuchten ADME-Gene und CNV-Bruchpunkten ergab, dass in Afroamerikanern 3,5%, in Weißen 10,9% und in Asiaten 7,6% der ADME-Gene mit CNVs assoziiert waren. Durchschnittlich wurden in Afroamerikanern 3,4 in Weißen 2,9 und in Asiaten 3,3 CNVs pro Patient gefunden (Insgesamt waren es 3,1 CNVs pro Donor). Gene mit CNVs, die in jeder Population in mindestens 10% der Proben gefunden wurden, waren hauptsächlich Gene der Phase II Gruppe, wie z.B. *GSTT1*, *GSTM1*, *UGT2B17* und *UGT2B28*. In allen Populationen wurden CNVs der Phase I Gene *CYP2A6* und *CYP2E1* gefunden. CNVs in Transportern (z.B.

ABCC1 und *6* und *SLC19A1*) sowie Genen der Gruppen ADME-Verwandte und Modifizierer (z.B. *COMT* und *AHRR*) wurden nur in Asiaten und Weißen beschrieben werden (Tabelle 7). In keiner Probe wurden *CYP2D6* CNVs detektiert.

In den Krebsproben wurden strukturelle Varianten in nahezu allen 340 ADME-Genen detektiert. Tumorspezifische Varianten, die in mehr als 60% aller Patienten auftraten, waren Deletionen im *NAT1* und *NAT2* Locus, sowie Duplikationen der Kernrezeptor und Transkriptionsfaktoren *NR1I3* (CAR), *NR5A2*, *ESRRG*, *ARNT*, *CRPK*, *ALDH9A1*.

Tabelle 7: CNV-Verteilung in ADME-Genen in gesunden Proben des TCGA-Projektes (n=348).

		Phasel	Phase II	Transporter	Modifizierer	ADME-verwandte
Schwarze oder Afroamerikaner	≥10%	CYP2A6, CYP2A7, CYP2E1, CYP4F12	GSTA1, GSTM1, GSTT1, GSTTP2, UGT2B15, UGT2B17, UGT2B28			
	1 < f < 10%		GSTA2			
Weiße	≥10%		GSTM1, GSTT1, GSTTP2, UGT2B15, UGT2B17, UGT2B28			
	1 ≤ f < 10%	ALDH2, ALDH3B2, CES1, CYP2A6, CYP2A7, CYP2E1, CYP4F12, EPHX1	GSTA1, GSTA2, GSTM2, GSTP1, UGT2B11	ABCC6, SLC16A3, SLC19A1	AHRR, ARSA	NUDT8
	<1%	CYP21A2, CYP4A11	GSTA5, SULT1A3, SULT2B1, UGT2B7	ABCA11P, SLC28A1		COMT, RNF40, STK19, VKORC1
Asiaten	≥10%	CYP2A6	CES1, GSTM1, GSTT1, GSTTP2, UGT2B17, UGT2B28			
	1 ≤ f < 10%	CYP2A7, CYP2E1, CYP4F2, EPHX1	GSTM2, GSTM4, UGT2B15	ABCC1, ABCC6		
	<1%	ALDH1B1, CYP2B6, CYP2B7P1, CYP4F12	SULT1A3, UGT2B11	ABCB7, SLC16A3, SLC19A1	AHRR	

3.1.3 CNV-Analyse in Leberdonoren der IKP148-Kohorte

3.1.3.1 CNV-Bestimmung über Abdeckungsdaten eines NGS-Projektes

Mit dem Kooperationspartner und der Sequenzierfirma CeGaT wurde ein ADME-Forschungspanel entwickelt und in DNA Proben der Leberdonoren (n=150) der IKP-148 Kohorte angewendet. Das Panel erlaubt die simultane Sequenzierung von 340 relevanten ADME-Genen. Über eine Analyse der sequenzierten DNA-Abschnitte (engl. *Reads*) und deren Abdeckung wurden CNVs in den ADME-Genen bestimmt. Die eigens hierfür adaptierte Methode wurde erfolgreich mit den X-chromosomalen Genen (*ABCB7*, *ABCD1*, *PGRMC1* und *TRPC5*) getestet. Wie erwartet wurde in männlichen Donoren ausschließlich ein Verhältnis der Abdeckung von 0,5, was einer Genkopienzahl von eins und in allen weiblichen Donoren ein Verhältnis um die Eins, was einer Kopienzahl von zwei entspricht, festgestellt (Abbildung 14).

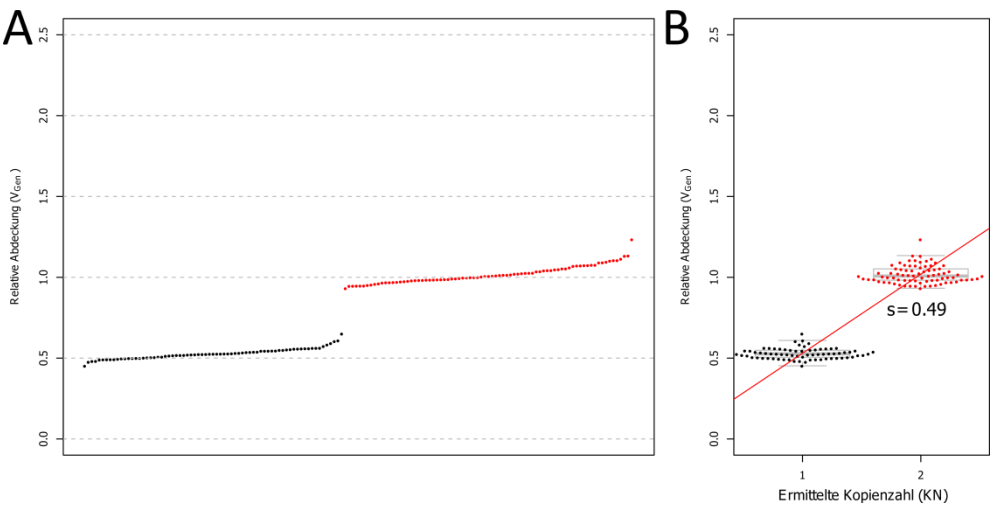


Abbildung 14: CNV-Analyse des X-chomosomalen *PGRMC1* Gens in 150 Leberdonoren (IKP148). **A** Jeweiliges Verhältnis der Abdeckung einer Probe gegen die erstellte Referenzprobe. **B** Gezeigt ist das Verhältnis aus A gegen die ermittelte Kopienzahl. Als rote Linie ist die Regressionsgerade gezeigt. Die Punkte sind nach Geschlecht des Leberspenders eingefärbt (Rot: weiblich; Schwarz: männlich).

Insgesamt wurden in den 150 Proben 561 CNVs ermittelt, 71% davon waren Deletionen und 29% Duplikationen. Durchschnittlich ergab dies 3,7 CNVs je Donor. Letztendlich waren 8% der Phase I, 18% der Phase II Gene, 4% der Transporter und ein Gen aus der Modifizierer-Gruppe CNV-assoziiert (Tabelle 8). Bestätigt wurden die bekannten CNVs in Genen der CYP-, UGT- und GST-Familie. Auch in diesem Datensatz zeigte sich das Gefälle der CNV-Frequenzen, beginnend bei Genen in der Gruppe der Transporter (selten; ≤1) über Phase I (Frequenzbereich von 5-10%) zu Phase II (häufig; >30%). In Modifizierern und ADME-Verwandten Genen kamen keine CNVs vor, mit Ausnahme des Gens *GPS2*, in dem in einem Leberdonor eine Kopienzahlvariation detektiert wurde (Tabelle 8). In der Phase I Gruppe waren Deletionen durchschnittlich seltener (3%) als Duplikationen (7%). Im Gegensatz dazu waren Deletionen in der Phase II Gruppe durchschnittlich häufiger (40%) als Duplikationen (13%). Die Analyse wurde anschließend auf Exonebene wiederholt. Dabei bedeutet Exonebene, dass nun, anstatt Exons zu einem Gen zusammenzufassen, für jeden Exonbereich die Kopienzahl erneut bestimmt wurde. In allen Proben und Genen, in denen CNVs mit der Analyse auf Genebene gefunden wurden, konnten dies in der Exon-Analyse nachvollzogen werden. Allerdings fanden sich Fälle, bei denen nicht alle Exons eines Gens eine veränderte Kopienzahl aufwiesen. Darüber hinaus wurden in bisher unauffälligen Proben einzelne CNV-assoziierte Exons detektiert. Am stärksten betroffen waren Fälle in *CYP2A6* und *CYP2D6*. Auf diese Ergebnisse wird im Teil 3.1.4 detaillierter eingegangen.

Tabelle 8: CNV-assoziierte ADME-Gene und deren Häufigkeit in den 150 Leberdonoren.

ADME-Gruppe	Gen	Deletion[1]	Duplikation[1]
Transporter	*ABCA2*	-	1%
	SLC2A4	-	1%
	SLC47A1	-	1%
Phase I	*CES1*	-	25%
	CYP21A2	1%	3%
	CYP2A6	2%	3%
	CYP2A7	7%	2%
	CYP2D6	7%	7%
	CYP2D7P1	1%	11%
	CYP2E1	1%	3%
Phase II	*GSTM1*	92%	-
	GSTT1	58%	-
	GSTT2B	-	5%
	SULT1A1	3%	34%
	UGT2B15	1%	1%
	UGT2B17	59%	-
	UGT2B28	29%	-
Modifizierer	*GPS2*	-	1%

[1]Genotyp Häufigkeit.

3.1.3.2 Bestätigung der CNV-Ergebnisse von ausgewählten ADME-Genen mit Hilfe einer CNV-Bestimmung per qPCR-Methode

Zur Validierung der oben beschriebenen Ergebnisse wurde für ausgewählte Gene die Kopienzahl mit TaqMan CNV-Assays und einer quantitativen PCR quantifiziert und anschließend mit den Ergebnissen der NGS-Methode verglichen.

Die CNV-Detektion der beiden Methoden war zu 98% identisch. Eine Diskrepanz der Bestimmung wurde für CNVs der Gene *CYP2D6*, *CYP2E1* und *SULT1A1* festgestellt. Dabei lieferte in einer Probe die exakte Quantifizierung duplizierter *CYP2E1* Kopien ein unterschiedliches Ergebnis (Abbildung 15), wobei zwei TaqMan CNV-Assays eine Kopienzahl von vier und die NGS-Methode eine Kopienzahl von drei berechnete. In drei Proben für *SULT1A1* war ebenfalls eine Diskrepanz der bestimmten Anzahl an dupli-

zierten Genkopien zwischen den beiden Methoden zu beobachten (Abbildung 15). Außerdem wurde in zwei Proben durch die qPCR-Analyse eine *SULT1A1* Duplikation vorhergesagt, die nicht in der NGS-Analyse detektiert wurde. Von der NGS-Analyse ermittelte Deletionen und Duplikationen des *CYP2D6* Gens wurden durch die qPCR-Methode in neun Proben nicht bestätigt. Alle weiteren Kopienzahlbestimmungen lieferten konsistente Ergebnisse zwischen den beiden Methoden. Die falsch-positiven oder negativen Ergebnisse für *CYP2D6* werden in 3.1.4 näher betrachtet und bewertet.

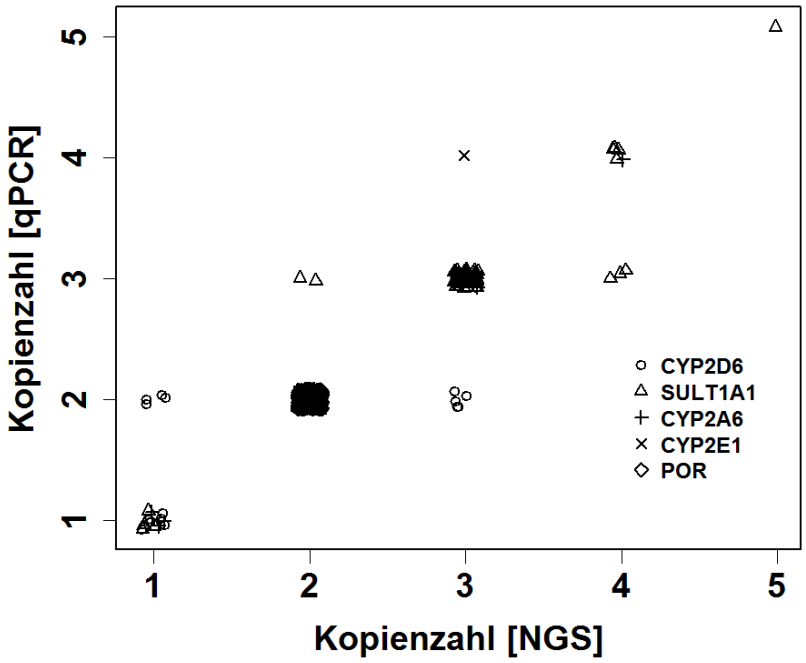


Abbildung 15: Vergleich der Kopienzahlbestimmung zwischen NGS- oder der qPCR-Methode mit TaqMan CNV-Assays für ausgewählte ADME-Gene.

3.1.4 Feinkartierung der *CYP2D6* und *CYP2A6* CNVs in Leberproben der Studie IKP148

Die durch die Auswertung der NGS-Daten der IKP148-Leberproben auf Exonebene gefundenen CNVs in den Genen *CYP2D6, CYP2D7P1, CYP2A6 und CYP2A7* wurden, auch auf Grund der unterschiedlichen Ergebnisse zur TaqMan Methode, im Folgenden detaillierter untersucht. Hierfür wurden neben der CNV-Information als zusätzliche Informationsquelle auch Genotypdaten aus verschiedenen vorausgegangenen Studien und Experimenten als Basis verwendet (Raimundo et al., 2000; Toscano et al., 2006; Zanger et al., 2001). Zusätzlich flossen Daten zu SNPs, die ebenfalls im ADME-Exonsequenzierungspanel erhoben wurden, in die Auswertung und Feinkartierung der Allele ein. Explizit wurde dabei die Variationsfrequenz (VAF) analysiert. Die VAF beschreibt den prozentualen Anteil an sequenzierten DNA Abschnitten, die eine bestimm-

te Variation tragen. Liegt ein SNP homozygot vor, beträgt die VAF= 0% oder VAF= 100%. Das heißt, alle Reads und somit beide homologen Chromosomen tragen an der SNP-Position das gleiche Nukleotid. Bei einem Träger mit einem heterozygoten SNPs beträgt im idealisierten Modell die VAF= 50%. Das bedeutet, dass theoretisch in einem deletierten DNA Locus nur VAF Werte um 0 oder 100% vorkommen können. In einer Person mit zwei Kopien eines DNA Bereichs können SNPs nur in drei Zuständen und damit drei VAF Clustern von 0%, 50% und 100% vorkommen. In einer Duplikation mit drei Kopien kann ein SNP in keiner, einer, zwei oder in drei von drei Kopien vorliegen. Somit sind Verhältnisse von 0%, 33%, 66% und 100% möglich (Abbildung 16).

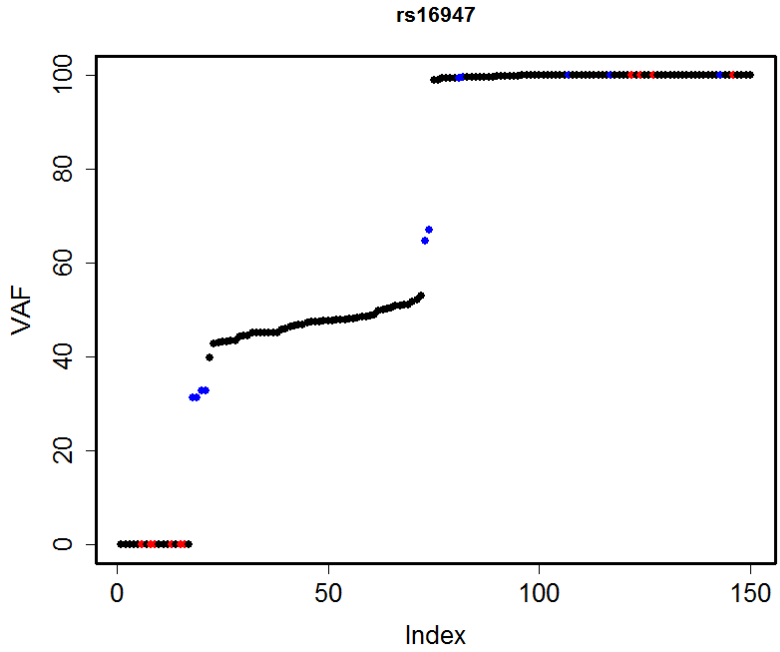


Abbildung 16: Zusammenhang von Variationsfrequenz und Kopienzahl in 150 Leberproben am Beispiel des SNPs rs16947 (*CYP2D6*2*). Die Kopienzahl ist farblich markiert. Dabei sind Proben mit einer Genkopie in Rot, Proben mit zwei Genkopien in Schwarz und Spender mit drei Kopien in Blau dargestellt.

3.1.4.1 Hybridallele der Gene *CYP2D6* und *CYP2D7P1*

Zu Beginn der CNV-Analyse für *CYP2D6* auf Exonebene wurde beobachtet, dass die mediane Abdeckung der Exons unterschiedlich und eine CNV-Bestimmung nicht in jedem Exon gleich zuverlässig war (Abbildung 17A). Außerdem ergab das Verhältnis der Abdeckung in Exon zwei, vier, sieben und acht in der Clusteranalyse keine eindeutig voneinander abgegrenzten CNV-Gruppen (Abbildung 17B). Deswegen konnte für diese Exons keine detaillierte und genaue Aussage zur Kopienzahl getroffen und keine definitive Auswertung gemacht werden. Die weitere Auswertung zeigte, dass Exon zwei die niedrigsten Abdeckungswerte aufwies (Abbildung 17A). Außerdem wurde eine

hohe Homologie der Exons zum Pseudogen *CYP2D7P1* bestätigt. Bei einem paarweisen Vergleich der Exon DNA Sequenzen zeigte sich, dass neben der insgesamt hohen Übereinstimmung der Exonsequenzen der beiden Gene, die durchschnittlich über 94% lag, Exon vier, sieben und acht nahezu identisch waren (Tabelle S3). So war nicht überraschend, dass die CNV-Bestimmung im *CYP2D7P1* in den Exons zwei, sieben und acht auch nicht exakt möglich war.

Für alle Proben die in der CNV-Analyse auf Genebene (Abschnitt 3.1.3) als *CYP2D6* CNV-Träger identifiziert wurden, konnte dies auch auf Exonebene nachvollzogen werden. Darüber hinaus wurde in 16 Proben eine Duplikation des Exons 1 und der Promoterregion detektiert (Abbildung 18; unterer Block).

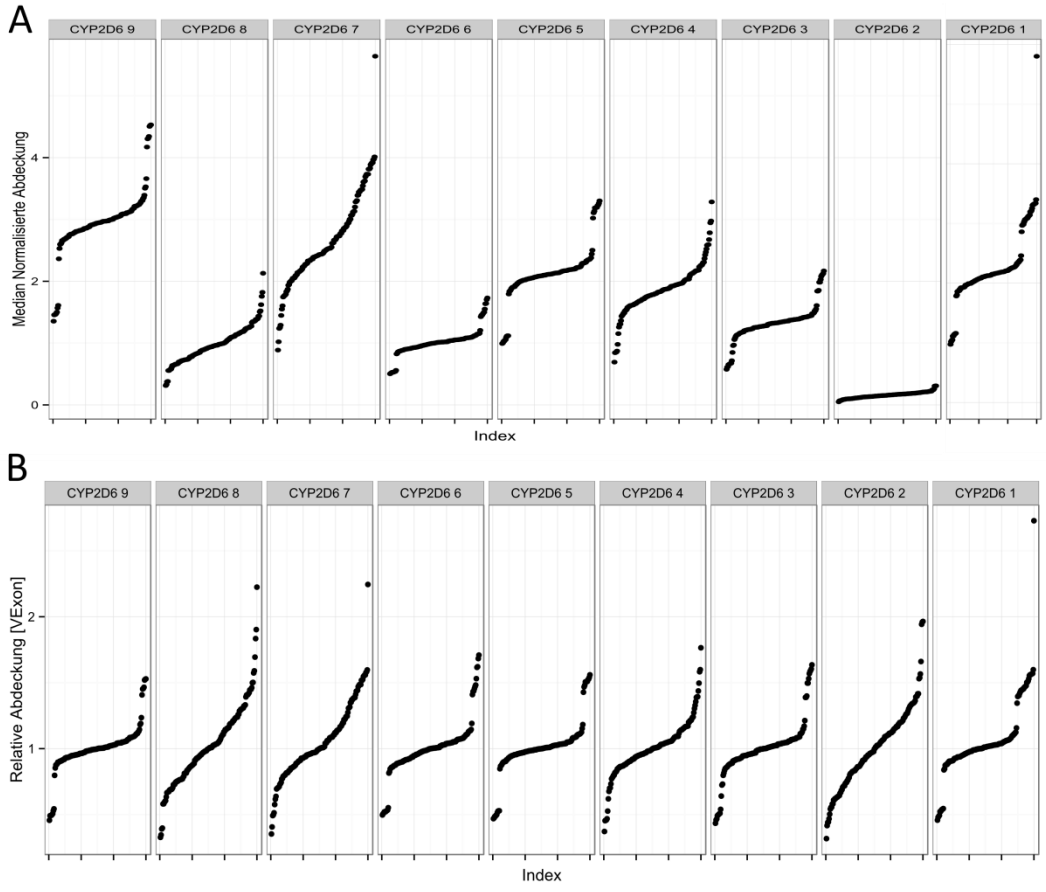


Abbildung 17: Übersicht der CNV-Analyse für jedes der neun Exons von *CYP2D6.* Dargestellt ist von rechts nach links Exon eins bis neun, was die Lage des Gens auf dem Minusstrang der DNA widerspiegelt. Exon eins umfasst zudem den Promoterbereich (2kb). **A** Normalisierte Abdeckungswerte **B** relative Abdeckung im Vergleich zur erstellten Referenzprobe für die 150 Leberproben (IKP148).

In sechs Proben wurde eine Deletion aller *CYP2D6* Exons detektiert, die auch mit dem TaqMan Assay (Hs00010001_cn), der in Exon neun bindet, gefunden wurde (Abbildung 18; erster Block). Diese Ergebnisse bestätigten das in diesen Proben vor-

liegende und bekannte heterozygote *CYP2D6*5* Allel. Des Weiteren wurde in vier Proben eine komplette Gendeletion (#166, #212) oder eine partielle, von Exon eins bis sechs reichende Deletion detektiert (#103, #165). Für diese vier Proben war bisher kein Deletionsallel bekannt und auch die Ergebnisse des TaqMan Assays zeigten keinen Kopienverlust. Da mit dem Assay eine Aussage zur Kopienzahl nur für Exon neun möglich ist, detektieren sowohl die Exonanalyse als auch die TaqMan Methode zwei Kopien für die Proben mit partieller Deletion in Exon neun. Der Vergleich der Ergebnisse zwischen NGS und TaqMan Methode für Exon neun erklärt die diskrepanten Ergebnisse in sieben von anfangs neun Proben (siehe Abbildung 15).

In den zwei Proben (#103, #165) mit partieller Deletion war zudem Exon acht und neun des Pseudogens *CYP2D7P1* deletiert. Weil ein CYP2D6-CYP2D7P1 Hybridallele bekannt ist, welches durch Genkonversion entsteht und zwischen Exon 7 und Intron 8 die Gensequenz wechselt, wurde für diese zwei Proben ebensolches angenommen. Diese Hypothese wurde durch die Analyse der SNP-Daten untermauert (Abbildung 19). Die zwei Proben trugen einen heterozygoten SNP (rs1058172) in Exon sieben, der typisch für das Hybridallel *CYP2D6*66* ist (Gaedigk et al., 2012). Zusätzlich wurde in den deletierten Exons eins bis sechs nur für eine Deletion typische homozygote SNPs beobachtet (Abbildung 19).

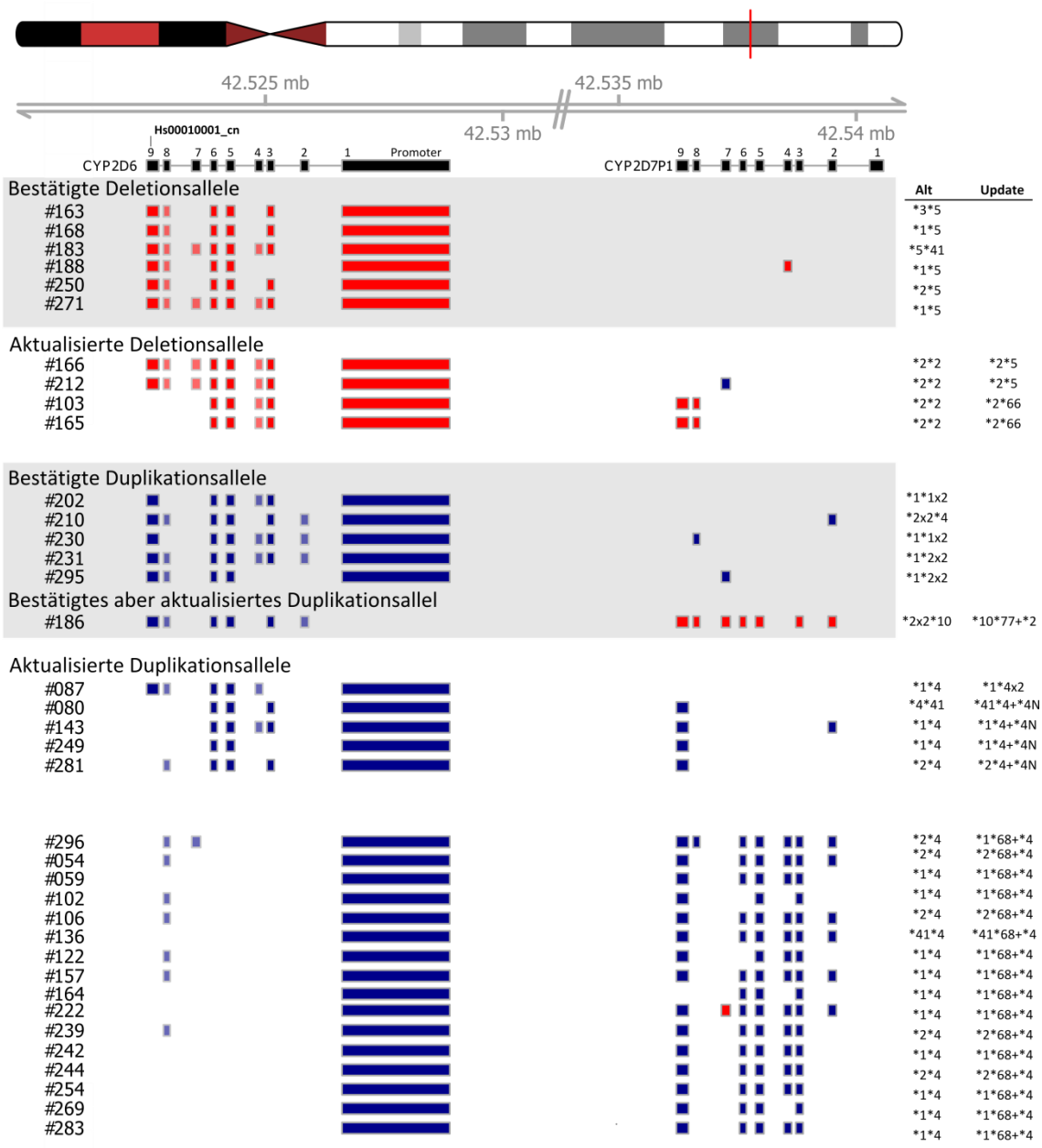


Abbildung 18: Kopienzahlunterschiede für jedes Exon von *CYP2D6* und *CYP2D7P1*. Von oben nach unten sind ein Ideogramm des Chromosoms 22, die Genomposition (GRCh37, hg19: Feb. 2009), die einzelnen Exons als schwarze Boxen und der TaqMan Assay zur Kopienzahlbestimmung mit qPCR in Exon neun dargestellt. Unterhalb der Exonregion befinden sich die individuell detektierten Varianten als Rechtecke. In Rot sind deletierte (Kopienzahl eins) und in Blau duplizierte Exons (Kopienzahl drei) markiert. Treffer in Exon zwei, vier, sieben und acht sind heller dargestellt, da die CNV-Detektion in diesen Bereichen nicht zuverlässig war. Die alten (Raimundo et al., 2000; Toscano et al., 2006; Zanger et al., 2001) und aktualisierten Allelinformationen sind tabellarisch auf der rechten Seite angegeben.

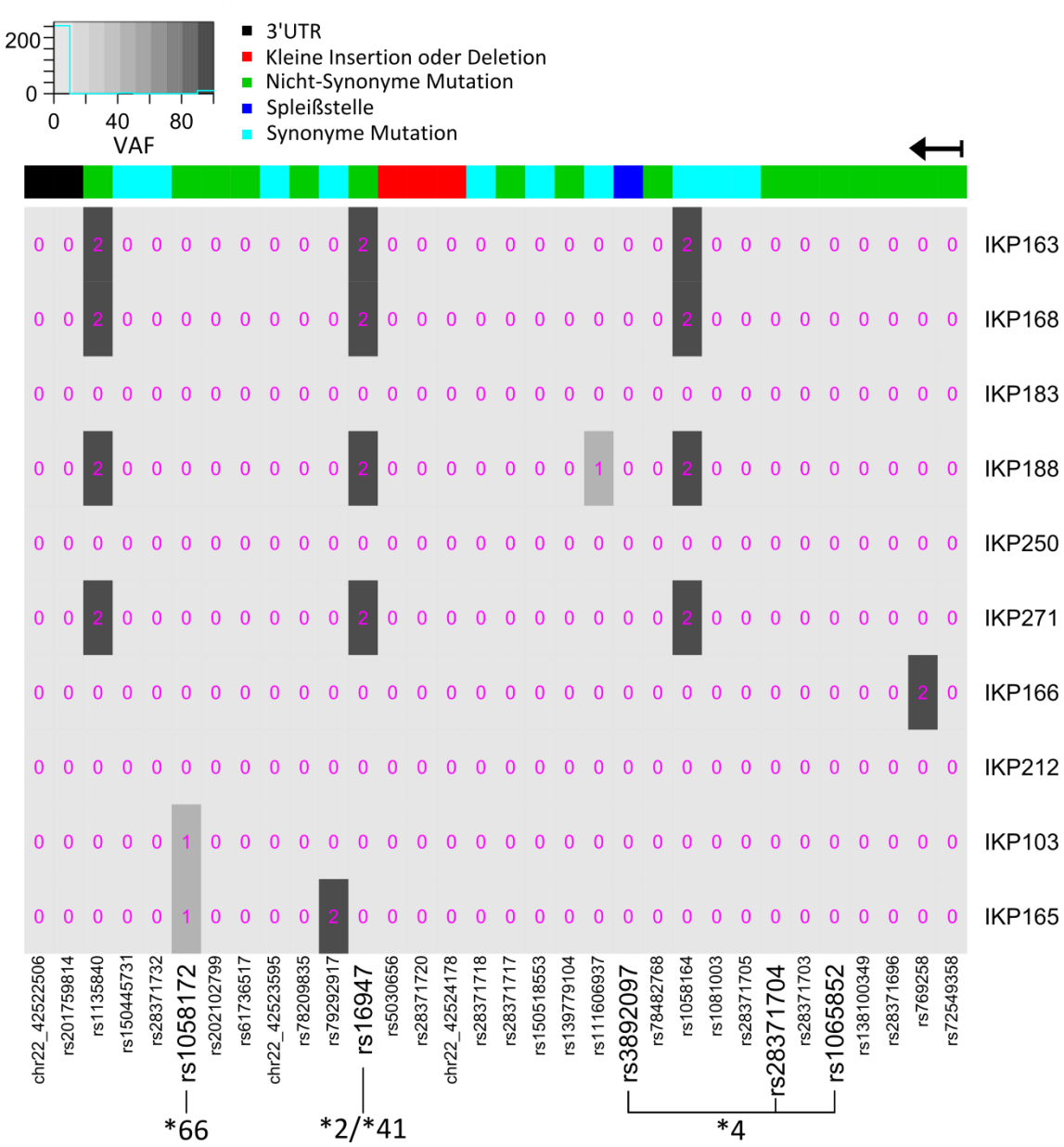


Abbildung 19: Variationsfrequenzen (VAF) und SNP Genotypen für Proben, die eine Deletion des *CYP2D6* Gens tragen. Die VAF Prozentwerte sind als Heatmap dargestellt. Ein helles Grau soll eine VAF von 0% und ein schwarzes Rechteck einen 100% Wert darstellen. Der dazugehörige Genotyp (als rosa Zahlen) ist numerisch kodiert (0=AA; 1= AB; 2= BB). Dabei entspricht eine Null oder eine Zwei dem jeweiligen homozygoten Zustand eines Allels. Eine Eins kodiert den heterozygoten Zustand. Der Pfeil zeigt die Leserichtung des Gens im Genom an.

In allen restlichen Proben mit Deletion, außer in Probe #188, wurde ebenfalls in jedem Exon ein Verlust von heterozygoten SNPs beobachtet (Abbildung 19). In Probe #188 trat ein heterozygoter SNP in Exon vier auf (rs111606937). Weil in dieser Probe genau das Exon vier des Pseudogens als deletiert detektiert wurde (Abbildung 18), musste davon ausgegangen werden, dass Reads des *CYP2D7P1* fälschlicherweise dem *CYP2D6* zugeordnet waren und somit einen falsch-positiven heterozygoten Zustand des SNPs vortäuschten. Tatsächlich war in dieser Probe an gleicher Stelle im Exon

vier des Pseudogens ein homozygoter SNP vorhanden, der die Sequenzsimilarität zum CYP2D6 steigerte (rs2267448).

Bekannte vollständige *CYP2D6* Genduplikationen (*CYP2D6*1x2* oder **2x2*) in sechs Leberproben wurden mit der NGS Exonlevel- und der VAF Clusteranalyse nachvollzogen (Abbildung 18 & 23). Dabei konnte für die Proben (#210, #231, #295, #186) mit einem duplizierten *CYP2D6*2* Allel (rs16947) eine VAF um 30% beobachtet werden. Da dieser SNP falsch annotiert war - das zur Zuordnung der 100bp Sequenzierungen verwendete Referenzgenom GRCh37 „trägt“ den SNP und damit sind Referenz und das alternative Allel vertauscht - bedeutet in diesem Fall eine VAF von 30%, dass nicht eine, sondern zwei von drei Genkopien den SNP trugen (Abbildung 20).

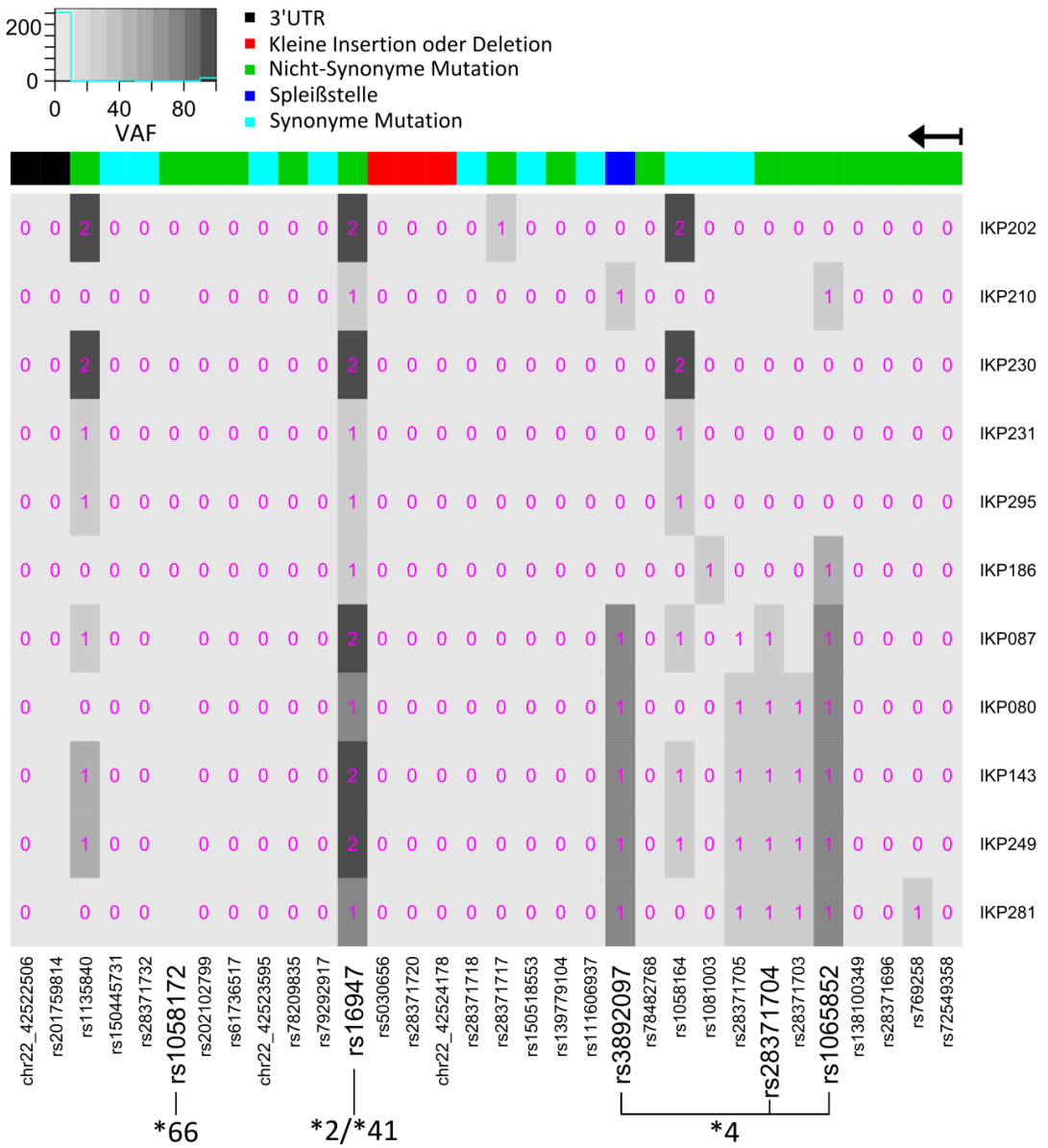


Abbildung 20: Variationsfrequenzen (VAF) und SNP-Genotypen in Proben die eine Duplikation des Gens *CYP2D6* tragen. Die VAF Prozentwerte sind als Heatmap dargestellt. Ein helles Grau soll eine VAF von 0% und ein schwarzes Rechteck einen 100% Wert darstellen. Der dazugehörige Genotyp (als rosa Zahlen) ist numerisch kodiert (0=AA; 1= AB; 2= BB). Dabei entspricht eine Null oder eine Zwei dem jeweiligen homozygoten Zustand eines Allels. Eine Eins einem heterozygoten Allel. Der Pfeil zeigt die Lage und Leserichtung des Gens im Genom an. Genotypen, die auf Grund der Filterkriterien nicht berücksichtigt werden konnten sind als leere Felder gezeigt.

In Probe #186 wurde eine Duplikation der Exons zwei bis neun festgestellt. Das erste Exon und der Promoterbereich waren nicht betroffen. Komplementär dazu waren die Exons zwei bis neun des Pseudogens *CYP2D7P1* deletiert. Zudem clusterten die VAF Werte von SNPs im Exon eins in dieser Probe um 0,50 oder 100%. Deswegen wurde für diese Probe der bisherige Genotyp *CYP2D6*2x2*10* zu dem Tandem Arrangement

*CYP2D6*10*77+*2* geändert. Der Kopf (Exon eins) des nicht funktionellen *77 Alleles besteht aus *CYP2D7P1* Sequenzen. Der restliche Teil aus *CYP2D6* Sequenzen.

In fünf Proben mit mindestens einem *4 Allel waren entweder alle Exons dupliziert (#087) oder alle bis auf Exon neun (#080, #143, #249, #281; Abbildung 17). Da in diesen Proben die VAF der im *4 Allel vorkommenden SNPs rs1065848 und rs3892097 um 60% clusterte (Abbildung 20), wurde eine Duplikation des *4 Allels (#087) oder Hybrid-*4 Allels, dessen Exon neun durch ein Rekombination aus *CYP2D7P1* Sequenzen besteht (#080, #143, #249, #281), bestimmt.

Die 16 Proben mit einer Duplikation des Promoter, 5'Bereichs und Exon eins trugen gleichzeitig eine Duplikation des Pseudogens ab Exon zwei (Abbildung 17) und wurden dem nicht-funktionellen *CYP2D6/2D7P1* Hybrid, das immer im Tandem mit einem *4 Allel zu finden ist, zugeordnet (*CYP2D6*4+*68*).

Insgesamt waren somit nur noch in drei Proben (#166, #212 und #087) unterschiedliche Ergebnisse zwischen NGS und TaqMan Methode für Exon neun festzustellen. Allerdings waren diese nicht weiter aufzulösen.

3.1.4.2 Hybridallele des Gens *CYP2A6* und dessen Pseudogens *CYP2A7*

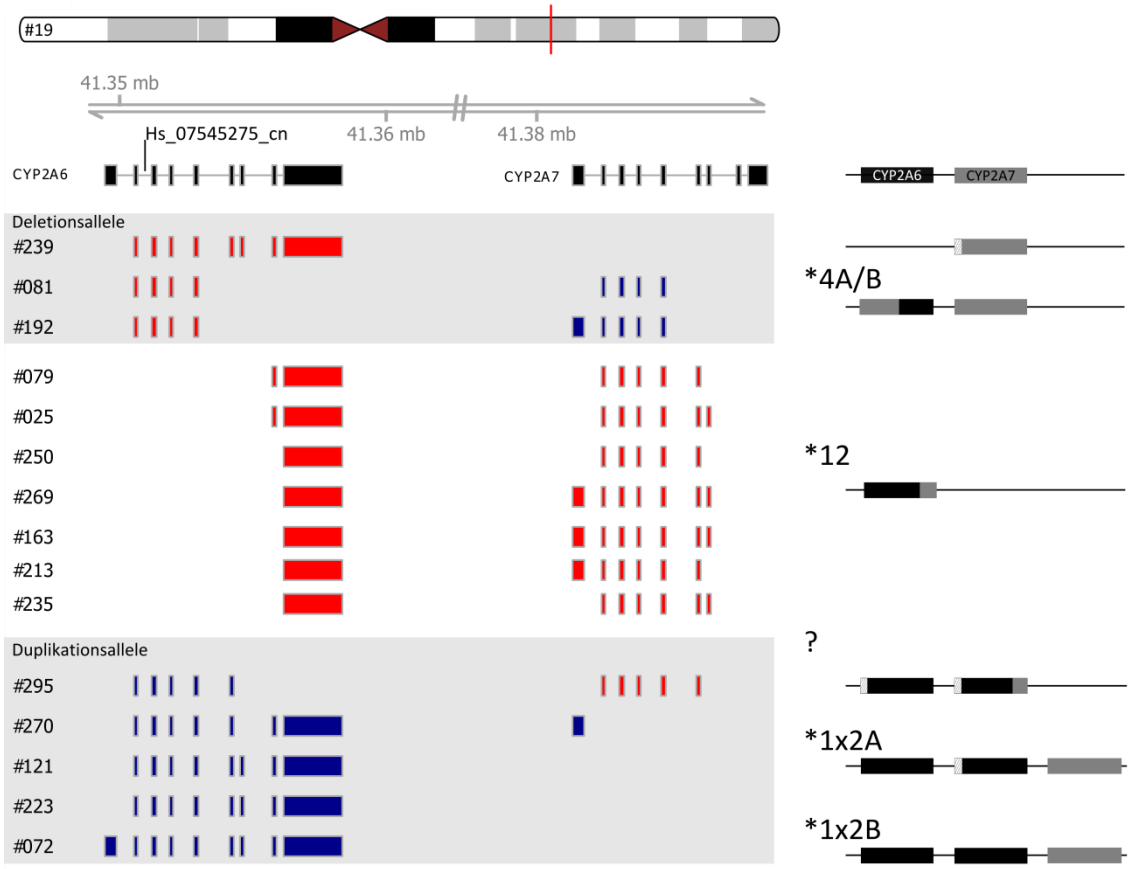


Abbildung 21: Unterschiede der Kopienzahl für jedes Exon von *CYP2A6* und *CYP2A7*. Von oben nach unten sind ein Ideogramm des Chromosoms 19, die Genomposition (GRCh37, hg19: Feb. 2009), die einzelnen Exons als schwarze Rechtecke und der TaqMan Assay zur Kopienzahlbestimmung in Intron sieben angezeigt. Unterhalb der Exonregion befinden sich die individuell detektierten Varianten als Rechtecke. In Rot sind deletierte (Kopienzahl eins) und in Blau duplizierte Exons dargestellt. Gleichförmige Ergebnisse der CNV-Detektionsmethoden NGS und qPCR sind grau hinterlegt. Auf der rechten Seite sind schematisch die angenommenen Hybridallele angezeigt. Unsichere Genkonversionen von *CYP2A6* zu *CYP2A7* sind als schraffierte Kasten gezeigt.

Wie bereits im *CYP2D6* Locus beobachtet wurde, unterschieden sich DNA Sequenzen zwischen *CYP2A6* und dessen Pseudogens *CYP2A7* in bestimmten Bereichen nur minimal, wie zum Beispiel dem Exon neun und der 3'UTR. Deswegen war eine Auswertung für *CYP2A6* in diesen Bereichen ebenfalls nicht exakt möglich. Es wurden drei Deletionen und fünf Duplikationen gefunden, die sowohl durch TaqMan (Intron 7) als auch durch die NGS-Analyse auf Exonlevel detektiert wurden (Abbildung 21; grau hinterlegt). Die drei Proben mit Deletion ließen sich durch die Anzahl an deletierten Exons unterschieden. Probe #239 trug eine bekannte Deletion der Exons ein bis acht (*CYP2A6*4*). In den weiteren Proben (#081, #192) waren die Exons fünf bis acht deletiert. Da genau diese Exons im *CYP2A7* dupliziert vorlagen, wurde davon ausgegan-

gen, dass in diesen beiden Proben ein bisher unbekanntes *CYP2A6-CYP2A7* Hybridallel vorliegt, dessen Bruchpunkt zwischen Exon vier und fünf liegen muss (Abbildung 21). In weiteren sieben Proben (#079- #235) wurde eine Deletion der *CYP2A6* Exons eins bis zwei und der *CYP2A7* Exons von drei bis acht oder neun beobachtet. Zusammengesetzt ergeben die nicht deletierten Exons das bekannte Hybridallel *CYP2A6*12* (Abbildung 21). In vier der fünf Leberproben, in denen eine *CYP2A6* Duplikation nachgewiesen wurde, konnte entweder das Duplikationsallel *CYP2A6*1x2A* (#270, #121, #223), das durch inäquales Crossing-over in einer Region in Intron 8 entsteht und somit ein *CYP2A7* Exon neun trägt, oder das Allel *CYP2A6*1x2B* (#072), dessen Bruchpunkt 5,2 bis 5,6kb hinter dem Stop Codon liegt und somit eine komplette *CYP2A6* Genduplikation darstellt, detektiert werden. Das Muster der deletierten und duplizierten *CYP2A6/7* Exons in Probe #295 war nicht exakt zu deuten. Es könnte sich aber um eine Kombination einer Duplikation *CYP2A6*1x2A* und einem *CYP2A6*12* Allel handeln. Der Vollständigkeit halber wurden weitere bereits beschriebene Allele mit Hilfe der SNP-Genotypen bestimmt. Wie in Abbildung 22 exemplarisch für die Proben mit veränderter Kopienzahl gezeigt, konnten funktionale Allele (*CYP2A6*17/*18*), das funktional eingeschränkte (*CYP2A6*21*) und das nicht-funktionelle Allel *CYP2A6*2* identifiziert werden.

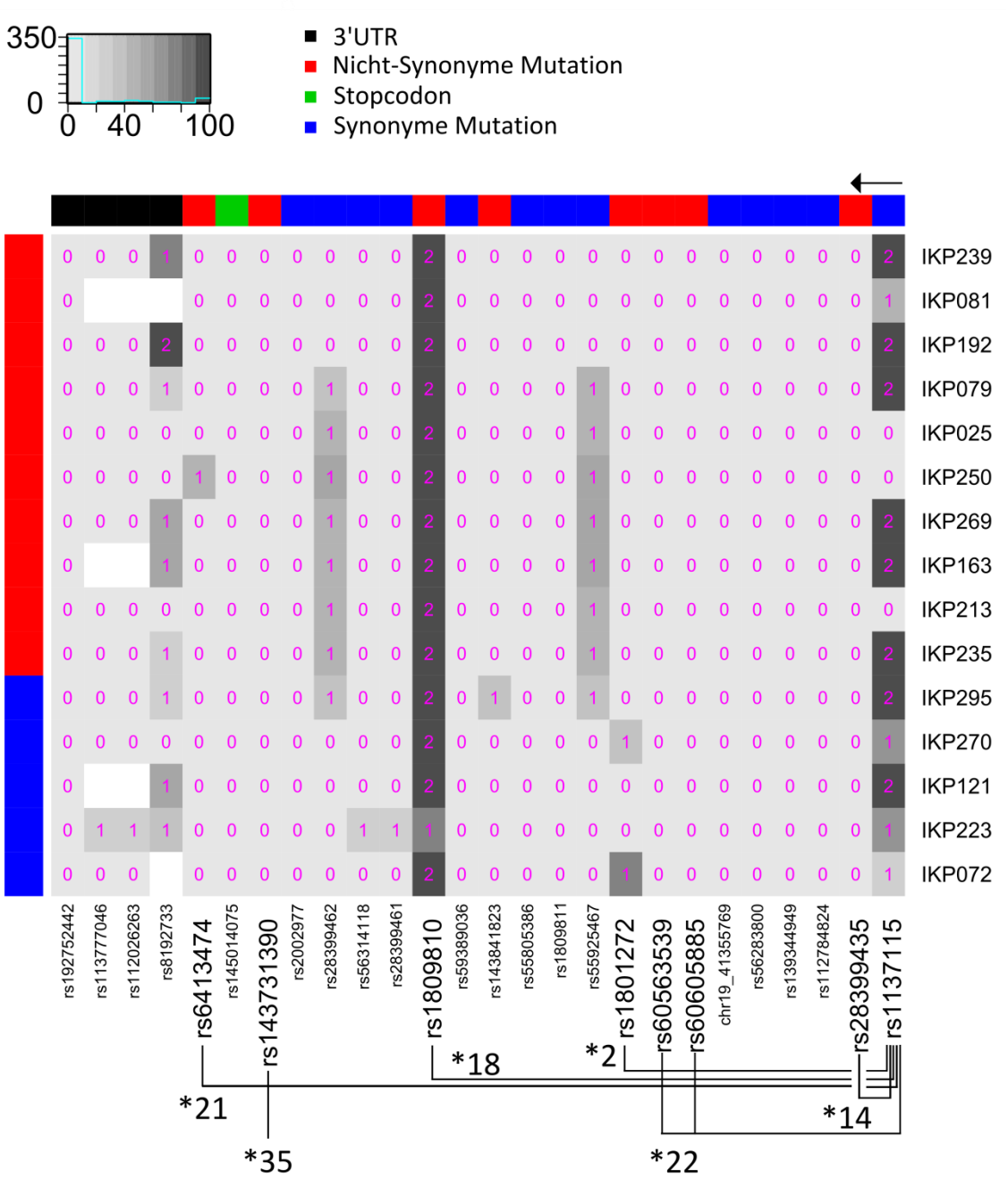


Abbildung 22: Variationsfrequenzen (VAF) und SNP-Genotypen in Proben mit veränderter Kopienzahl. Die VAF Prozentwerte sind als Heatmap dargestellt. Ein helles Grau soll eine VAF von 0% und ein schwarzes Rechteck einen 100% Wert darstellen. Der dazugehörige Genotyp (als rosa Zahlen) ist numerisch kodiert (0=AA; 1= AB; 2= BB). Dabei entspricht eine Null oder eine Zwei dem jeweiligen homozygoten Zustand eines Allels. Eine Eins einem heterozygoten Allel. Der Pfeil zeigt die Lage und Leserichtung des Gens im Genom an. Im unteren Bereich sind bekannte *CYP2A6* Allele (http://www.cypalleles.ki.se; August 2015) markiert. Die Spaltenmarkierung auf der linken Seite deutet die Kopienzahl der jeweiligen Probe an, wobei blau einen Gewinn und rot einen Verlust von *CYP2A6* Exons in der jeweiligen Probe anzeigt.

3.1.5 Vergleich des CNV-Vorkommens im IKP148-Datensatz mit dem der DGV-Datenbank und den Leberproben des TCGA- Projektes

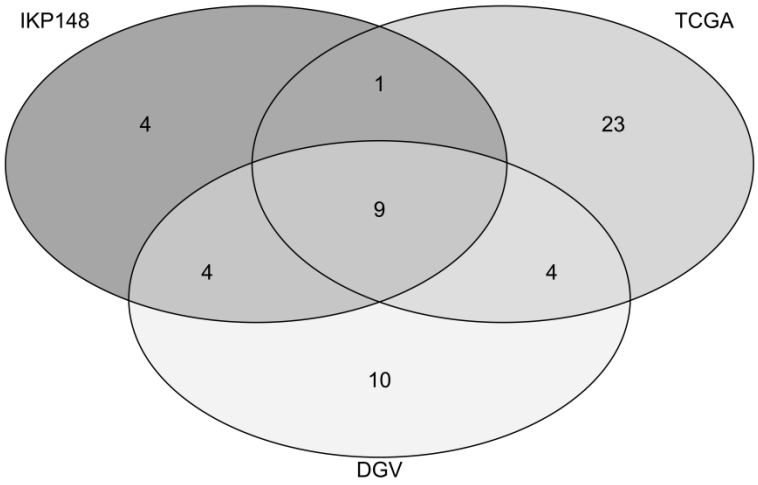


Abbildung 23: Venndiagramm zur Veranschaulichung der Ergebnisse der CNV-Bestimmung in ADME-Genen in den Kohorten DGV, TCGA und IKP148. Die Zahlen geben die Anzahl der CNV-assoziierten Gene in den jeweiligen Sektionen an. Zur besseren Vergleichbarkeit wurden die Ergebnisse auf Daten beschränkt, die ausschließlich in Proben mit europäischer Herkunft erstellt wurden.

Der Vergleich der beobachteten CNVs in Proben mit Europäischer Abstammung zeigte, dass Personen Europäischer Abstammung durchschnittlich 3,6 ±0,6 CNVs in den untersuchten ADME-Genen (n=340) trugen. Probanden asiatischer Herkunft trugen 3,8 ±0,5 CNVs und in Proben mit afrikanischen Wurzeln 4,0 ±0,6 CNVs pro Donor auftraten. In allen Kohorten waren CNVs in Genen der GST und UGT Familie am häufigsten, gefolgt von CNVs in der CYP450 Genfamilie. CNVs in Transportern waren selten. Die in den drei Kohorten ermittelten CNV-Frequenzen in Europäischen Proben waren zu publizierten Frequenzen vergleichbar (Tabelle 9).

In neun ADME-Genen wurden CNV in allen drei Kohorten gefunden (Abbildung 23). Darunter befanden sich die CNVs von Genen der Phase I und II wie *CYP2A6* und dessen Pseudogen 2A7, *CYP2E1*, *CYP21A2* sowie Treffer in Genen der GST- (GSM1, GSTT1) und UGT-Familie (*UGT2B15*, *UGT2B17*, *UGT2B28*). Nur in der IKP148-Studie wurden CNVs in den Transporten *SLC2A4*, *SLC47A1*, *ABCA2* und *GPS2* detektiert. CNVs, die in den Leberdonoren (IKP148) und jeweils in DGV oder TCGA gefunden wurden, waren *SULT1A1*, *GSTT2B* und der *CYP2D6* Locus inklusive dessen Pseudogen *CYP2D7P1*, sowie *SULT1A3* und *CES1*. Der große Anteil der individuellen CNV-Treffer des TCGA-Projektes umfasste ausschließlich selten auftretende (Frequenz <10%) CNVs in Transportern und Genen der GST Familie. Ausschließlich in HapMap-Proben wurden CNVs im FMO-Locus sowie in diversen Transportern gefunden.

Tabelle 9. Vorkommen der IKP148-CNVs in Proben mit europäischer Abstammung im Vergleich zu den anderen Kohorten und publizierten Daten (Genotyp-Frequenz).

ADME-Gruppe	Gen	Deletion				Duplikation				Referenz
		IKP148 (n=150)	TCGA (n=156)	DGV (n=184)	Publiziert	IKP148 (n=150)	TCGA (n=156)	DGV (n=184)	Publiziert	
Transporter	*ABCA2*	-	-	-	-	1%	-	-	-	
	SLC2A4	-	-	-	-	1%	-	-	-	
	SLC47A1	-	-	-	-	1%	-	-	-	
Phase I	*CES1*	-	-	-	-	25%	8%	-	27.8%	Fukami et al. (2008)
	CYP21A2	1%	1%	2%	-	3%	-	1%	7%	Parajes et al. (2008)
	CYP2A6	2%	1%	1%	1%	3%	1%	-	3%	Haberl et al. (2005)
	CYP2A7	7%	-	13%	-	2%	1%	1%	-	
	CYP2D6	7%	-	1%	6%	7%	-	1%	9%	Zanger et al. (2008)
	CYP2D7P1	1%	-	1%	-	11%	-	1%	-	
	CYP2E1	1%	-	-	-	3%	8%	3%	6%	Martis et al. (2012)
Phase II	*GSTM1*	92%	59%	96%	93%	-	-	-	-	Rose-Zierilli et al. (2009)
	GSTT1	58%	39%	93%	68%	-	-	-	-	Rose-Zierilli et al. (2009)
	GSTT2B	-	-	76%	-	5%	-	8%	-	
	SULT1A1	3%	-	4%	6%	34%	-	1%	27%	Hebbring et al. (2007)
	UGT2B15	1%	1%	-	1%	1%	13%	7%	-	Gaedigk et al. (2012)
	UGT2B17	59%	58%	64%	56%	-	-	-	0.2%	Gaedigk et al. (2012)
	UGT2B28	29%	37%	29%	25%	-	-	-	-	Menard et al. (2009)
Modifizierer	*GPS2*	-	-	-		1%	-	-	-	

3.2 Einfluss von CNVs auf die Genexpression von ADME-Genen in humanen Proben

Die Expression der ADME-Gene in den Leberproben der IKP148-Studie und LCLs der HapMap-Proben wurde aus Mikrochip Experimenten extrahiert. Für die Leberproben des TCGA-Projektes wurden genomweite RNA Sequenzierdaten ausgewertet. Die Genexpressionen wurden anschließend jeweils gegen die im Abschnitt 3.1 detektierten CNVs korreliert.

3.2.1 Assoziationsanalyse zwischen CNVs und Mikrochip Genexpressionsdaten aus LCLs der HapMap-Proben

Die Assoziationsanalyse zwischen CNVs und Expressionsdaten in 270 HapMap-Proben und 15 Genen, für die sowohl CNV als auch Expressionswerte vorlagen, zeigte, dass die Deletionen der Gene *GSTM1* und *GSTT1* den signifikantesten Einfluss (p =1,4E-15 und p=1,8E-14) auf die Genexpression der beiden Gene hatten (Abbildung 26). Des Weiteren wurde für *UGT2B17* CNVs (p= 2,6E-13) ein signifikanter Einfluss beobachtet. Für alle weiteren CNVs von Genen wie *CYP2E1* und *CYP4F2*, *FMO4*, *GSTM2* und *SULT1A1* wurde kein Einfluss auf die jeweilige Genexpression gezeigt.

3.2.2 Assoziationsanalyse von CNVs mit RNA-Seq Genexpressionsdaten des TCGA-Datensatzes

Gesundes Gewebe

Im TCGA-Datensatz lagen Genexpressionswerte von 50 gepaarten Proben vor. Eine Hauptkomponentenanalyse (PCA) zeigte, dass sich das Expressionsschema im Tumorgewebe von dem im gesunden Lebergewebe unterscheidet (Abbildung 24A). Allerdings fanden sich zwei Tumorbiopsien, die ein ähnliches Expressionsmuster wie gesunde Proben aufwiesen und wurden aus der späteren Analyse ausgeschlossen. Mit einer differentiellen Expressionsanalyse wurden Expressionsunterschiede zwischen Normal- und Tumorgewebe beobachtet (biologischer Variationskoeffizient (BCV= 0,79). Im Tumor höher exprimierte Gene wurden in einer GO Analyse dem zellulären Protein Metabolismus, der Antwort auf chemische Reize, Phosphorsäureester Biosynthese und Metabolismus und dem Zellzyklus zugeordnet. Eine geringere Expression im Tumor war für Gene der Akut-Phase-Antwort, des Steroid Metabolismus, des Carbonsäure Metabolismus und ebenfalls der Phosphorsäureester Biosynthese zu beobachten. Da die 50 Proben aus unterschiedlichen Populationen (Asien; n=5, Afrika; n=7, Europa; n=34) stammten, wurde zudem analysiert, ob die Abstammung der Probe ei-

nen Einfluss auf die Expression im gesunden Gewebe hat. Die Analyse zeigte moderate Unterschiede mit einem biologischen Variationskoeffizient von 0,46 (Abbildung 24C). Die Expression von ADME-Genen, unter anderen die von *CYP2A6*, *CYP2C19*, *CYP2D6* und *CYP2E1* sowie den Transportern *SLC29A4* und *SLCO1A1* war in Proben mit afrikanischer und asiatischer Abstammung im Vergleich zu Proben europäischer Herkunft unterschiedlich.

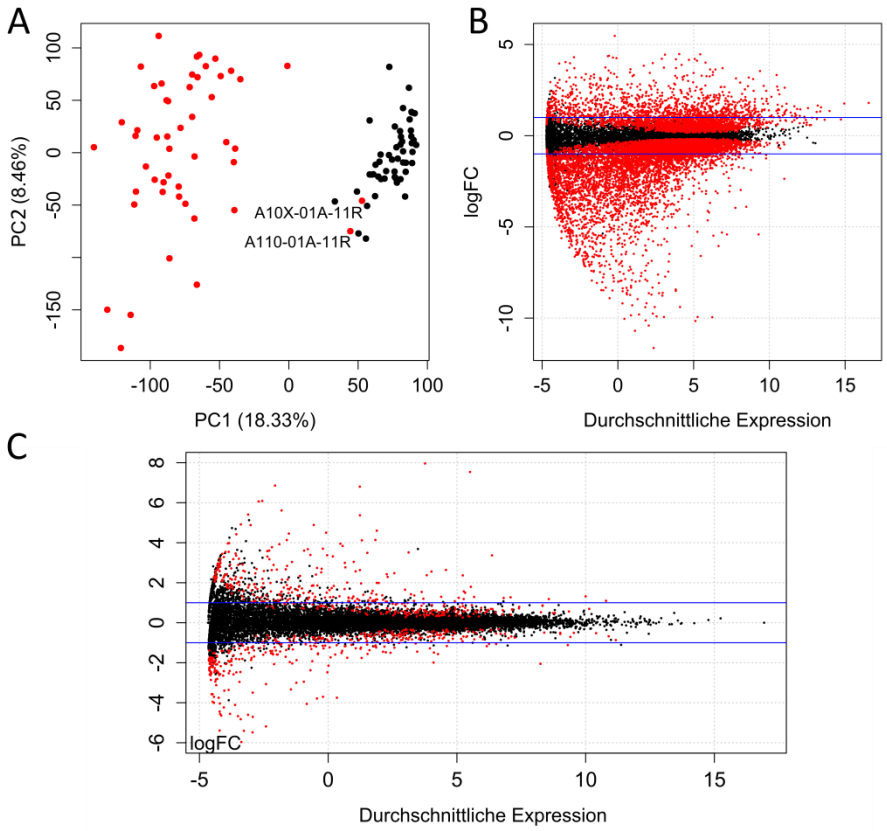


Abbildung 24 **A** Hauptkomponentenanalyse der Genexpressionswerte (log2) von gesundem (Schwarz) und Tumorgewebe (Rot). Der Prozentwert beschreibt den Beitrag zur Gesamtvariabilität. **B** Differentielle Genexpressionsanalyse zwischen der Expression in Tumor und Normalgewebe. Negative Werte beschreiben eine im Tumor erhöhte Expression. **C** Expressionsunterschied zwischen Proben mit asiatischer und afrikanischer Herkunft und Proben europäischer Abstammung. In Rot sind signifikante, nach Benjamini Hochberg adjustierte Unterschiede hervorgehoben. Die horizontalen blauen Linien markieren einen 2-fachen Expressionsunterschied.

Für 47 gesunde Gewebeproben und 12 der 48 CNV-assoziierten ADME-Gene lagen im TCGA-Datensatz sowohl Genexpressionsdaten als auch CNV Genotypen und die Information der Abstammung vor. Die Korrelation zwischen CNV und mRNA wurde mit einem multivariaten Modell, das die Abstammung der Proben berücksichtigt, berechnet. Eine statistisch signifikante Assoziation wurde für *GSTM1* (p= 1.2E-4) und *GSTT1*

(p= 9,1E-6) gefunden. Alle weiteren CNVs der *Gene CYP2A6, CES1, CYP2E1, CYP4F12, UGT2B15, UGT2B28, GSTA1, GSTA2*, und *NUDT8* waren nicht zur Expression assoziiert (Abbildung 26).

Tumorgewebe

Im Tumorgewebe konnte die Assoziationanalyse mit insgesamt 320 Spendern und 339 ADME-Genen wiederholt werden. Nach einer p-Wert Adjustierung mit der Bonferroni-Methode wurde in 35% der 340 untersuchten Gene eine signifikante Korrelation ($p<0.05$) zwischen CNV und Expression festgestellt (Abbildung 25). Die signifikantesten Einflüsse wurden, wie auch im gesunden Gewebe, für *GSTM1* und *GSTT1* CNVs (log2 fc=4,4 und 5,1) und von CNVs der Ubiquitin Ligase RNF40 (log2 fc=0,85) gefunden. Die Top zehn Assoziationen wurden durch CNVs in den Genen *ABCF1*, *ABCD4*, *CARM1*, *GPS2*, *ABCF2*, und *CHURC1* vervollständigt (Abbildung 25 und Abbildung A1).

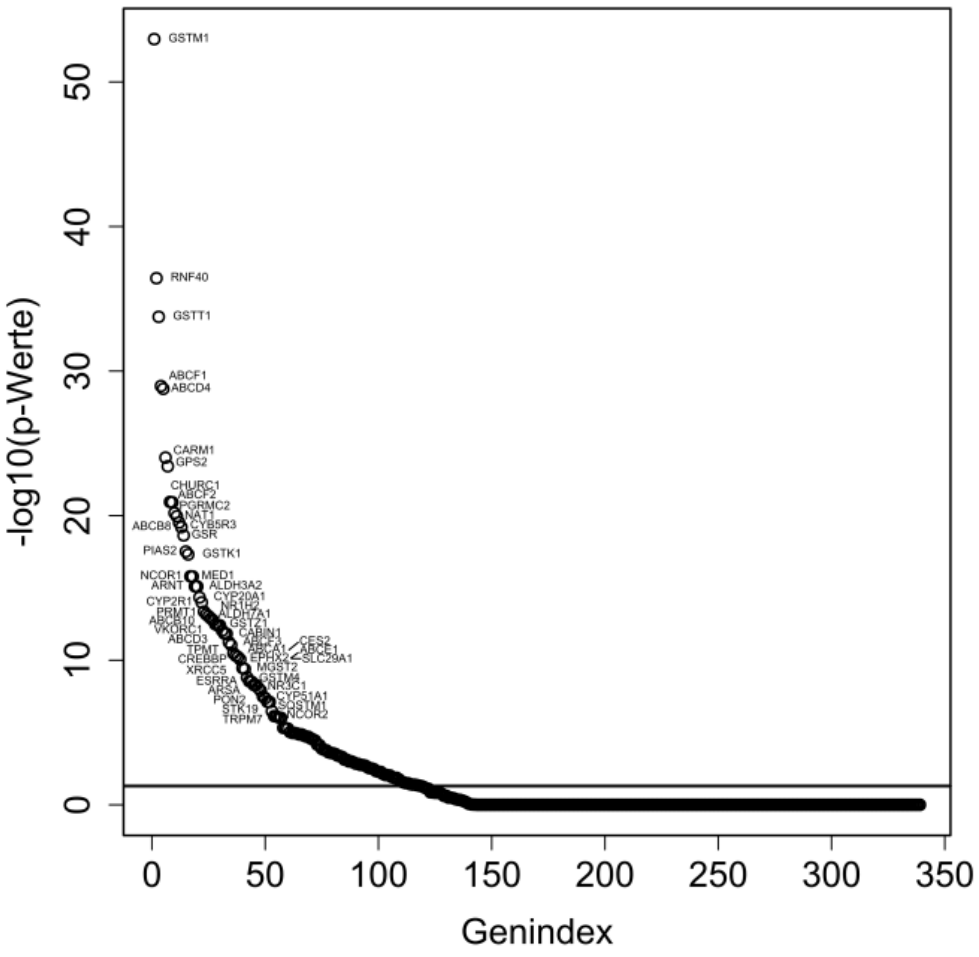


Abbildung 25: Ergebnis der Assoziationsanalyse zwischen CNVs und mRNA Expression in ADME-Genen in Tumorlebergewebe. Gezeigt sind nach Benjamini-Hochberg korrigierte p-Werte. Die horizontale Linie beschreibt das Signifikanzniveau bei 0,05.

3.2.3 Assoziationsanalyse von CNVs und Genexpressionsdaten in humanen Leberproben der Studie IKP148

Für die Assoziationsbestimmung wurde die aus der Abdeckungsanalyse auf Genebene gewonnene relative Kopienzahl verwendet. Für alle Gene, die CNVs aufwiesen (außer *GSTT2B*) waren Expressionswerte eines Mikrochipexperimentes verfügbar (Schröder et al., 2013). Die Assoziationsanalyse ergab in sieben der 17 untersuchten CNV-assoziierten Gene einen signifikanten Zusammenhang zwischen Kopienzahl und Expression (Abbildung 26, Kruskal-Wallis Test). Darunter befand sich die Expression von *GSTM1*, *GSTT2*, *UGT2B17*, *SULT1A1* und *CYP2D6* und dessen Pseudogens. Die Expression der weiteren Gene war nicht mit der Kopienzahl assoziiert (Abbildung 26). Für CNVs in der Transporter Genfamilie war die niedrige CNV-Frequenz und somit der geringe Probenumfang von nur einer Probe für den hohen p-Wert verantwortlich. Nach Korrektur für multiples Testen waren nur die Assoziationen der *GSTM, GSTT1* und *SULT1A1* CNVs signifikant.

3.2.4 Vergleich und Zusammenfassung der Ergebnisse der Assoziationsanalyse

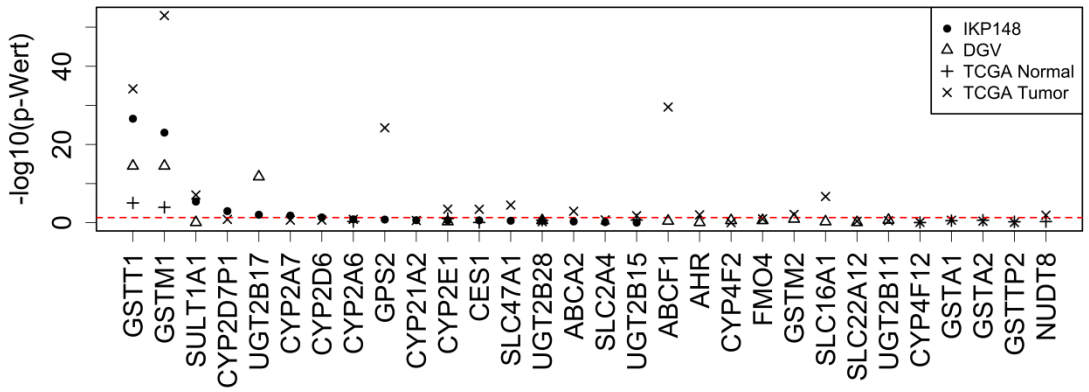


Abbildung 26 Vergleich der nach Benjamini-Hochberg korrigierten p-Werte aus den Assoziationsanalysen zwischen mRNA Expression und Kopienzahl in den Kohorten DGV, TCGA und IKP148. Es lagen nicht für jedes Gen Expressionswerte in allen Datensätzen vor.

Der Vergleich der berechneten Assoziationen in den Kohorten DGV und TCGA und der Analyse in den Leberproben der Studie IKP148 zeigte deutlich, dass die häufigen CNVs in der GST- Familie in jeder Kohorte am stärksten zur Expression assoziiert waren. Ein Vergleich der Proben mit europäischer Abstammung verdeutlichte diesen Einfluss (Abbildung 27 & Tabelle 10). In HapMap-Proben mit Verlust von *GSTT1* oder *GSTM1* waren zwei Expressionscluster zu beobachten, die mit hoher Wahrscheinlichkeit einer homozygoten und heterozygoten Deletion zugeordnet werden können. Diese

Information stand in DGV aber nicht zur Verfügung (Abbildung 27; erste und zweite Reihe, zweite Spalte).

Schwächere positive Assoziationen, die aber nicht in allen Kohorten zu finden waren, wurden für *SULT1A1*, *CYP2D6*, *CYP2A6* und deren Pseudogene *CYP2A7* und *CYP2D6P1* beobachtet (Abbildung 26). *SULT1A1* CNVs waren in den Leberproben (IKP148) und in europäischen TCGA-Tumorproben signifikant zur mRNA Expression assoziiert (Abbildung 27). In HapMap-Proben und in den gesunden TCGA-Leberproben wurden entweder keine Duplikationen oder überhaupt keine CNVs im *SULT1A1* Locus detektiert. Gleiches galt für CNVs in *CYP2A6* und *CYP2D6*, welche in den IKP148 Leberproben signifikant mit der mRNA Expression korrelierten (Abbildung 27; vierte und fünfte Reihe). Interessanterweise war die Assoziation für *CYP2A6* CNVs im Tumor umgekehrt. Es lag eine höhere Expression in Proben mit Deletion und eine erniedrigte mRNA Expression in Proben mit Duplikation vor.

Tabelle 10: Vergleich der Assoziation von CNV und mRNA Expression zwischen den in den Leberproben (IKP148) gefundenen CNVs und den anderen Kohorten in Proben mit europäischer Abstammung.

CNV in Gen	**IKP148[1]**	**DGV(HapMap)[1]**	**TCGA (N)[1]**	**TCGA (T)[1]**
GSTM1	1,1E-24	0,0004	0,0002	1,6E-16
GSTT1	1,5E-28	1,7E-13	5,6E-06	1,3E-14
SULT1A1	7,2E-07	n.s.		0,0002
CYP2A6	0,03		n.s.	0,006
CYP2D6	0,02			0,006
UGT2B17	0,003	1,0E-07		
CES1	n.s.		n.s.	4,4E-07
CYP2E1	n.s.	n.s.	n.s.	n.s.
CYP21A2	n.s.			n.s.
UGT2B15	n.s.		n.s.	0,001
UGT2B28	n.s.	n.s.	n.s.	n.s.

[1]Die dargestellten unadjustierten p-Werte wurden mit einem univariaten Kruskal-Wallis Test berechnet. Das Signifikanzniveau lag bei $p<0.05$. Standen keine Expressionswerte oder CNV-Daten zur Verfügung, fehlt der Eintrag.

UGT2B17 CNVs waren in Leberproben (IKP148) und in LCLs der HapMap-Proben zur mRNA Expression assoziiert (Abbildung 16). Allerdings fiel bei der Auswertung der Boxplotdiagramme in Abbildung 27 auf, dass in den IKP148-Leberproben trotz einer

homozygoten Deletion des *UGT2B17* Gens eine Expression gemessen wurde. Die erneute Genomlokalisation der Probensequenzen des Mikrochips zeigte, dass alle auf dem Mikrochip angegebenen Sequenzen für *UGT2B17* auch zu anderen homologen UGT-Formen, wie *UGT2B4*, *UGT2B10* und *UGT2B15* passen und die gemessene Expression nicht unbedingt UGT2B17 spezifisch war. Es wurde keine Assoziation zwischen den *UGT2B17* CNVs und der mRNA Expression von UGT2B4, UGT2B10 und UGT2B15 beobachtet. Im TCGA-Datensatz wurden zwar CNVs detektiert, allerdings waren keine RNA Sequenzierungsdaten und mRNA Expressionswerte für UGT2B17 verfügbar. In HapMap-Proben mit einem Verlust von *UGT2B17* konnten wieder zwei Probencluster identifiziert werden (Abbildung 27). Eine Gruppe wies dabei log2-Expressionswerte von sechs auf, was bei dem verwendeten Mikrochip auf eine minimale bis keine Expression hindeutet und vermuten ließ, dass diese Gruppe eine homozygote *UGT2B17* Deletion trug. Die UGT2B17 mRNA Expression in HapMap-Proben mit einem Gengewinn war im Vergleich zu Proben mit zwei Kopien nicht erhöht.

Für CNVs in *CES1* und *CYP2E1* wurde in der populationsabhängigen multivariaten (Abbildung 26) sowie in der univariaten Analyse (Abbildung 28) in Proben mit europäischem Ursprung keine Assoziation zur mRNA Expression gefunden. Der in Lebertumorproben beobachtete schwache signifikante Einfluss auf die CES1 Expression (Tabelle 10) basierte auf den in gesunden Proben nur sehr selten vorkommenden Deletionen. So war die CES1 Expression in Trägern einer Deletion im Vergleich zu Proben mit normaler Kopienzahl signifikant 1,4-fach (log2) erniedrigt (Wilcoxon-Vorzeichen-Rang-Test; p= 1,4E-07). Es konnte kein signifikanter Expressionsunterschied zwischen zwei und drei *CES1* Kopien festgestellt werden (Abbildung 28; TCGA-Tumorproben).

Ähnliches war für das Gen *CYP21A2* (Abbildung 28) zu beobachten. In den IKP148-Leberproben hatten Träger einer Deletion des Gens im Vergleich zu Proben mit normaler Kopienzahl eine tendenziell erniedrigte mRNA Expression (n.s.). In Trägern einer Duplikation war keine Steigerung der Expression im Vergleich zu Trägern von zwei Genkopien zu erkennen (Abbildung 28).

CNVs des Gens *UGT2B28* und die seltenen CNVs des Gens *UGT2B15* zeigten keinen Einfluss auf die jeweilige mRNA Expression. Wie oben bereits für *UGT2B17* beobachtet, wurde in IKP148-Leberproben mit einer homozygoten Deletion trotz eines Nullallels eine mRNA Expression gemessen (Abbildung 27).

Die weiteren CNV-assoziierten Gene, die exklusiv nur in den HapMap- oder TCGA-Proben gefunden wurden, zeigten größtenteils eine Gendosis-unabhängige Expression

und wurden aus verschiedenen Gründen, wie zu geringe Probenanzahl mit CNV, zu niedrige Genexpression und unklare bzw. falsch-positive CNV-Treffer nicht detaillierter betrachtet (Abbildung 26).

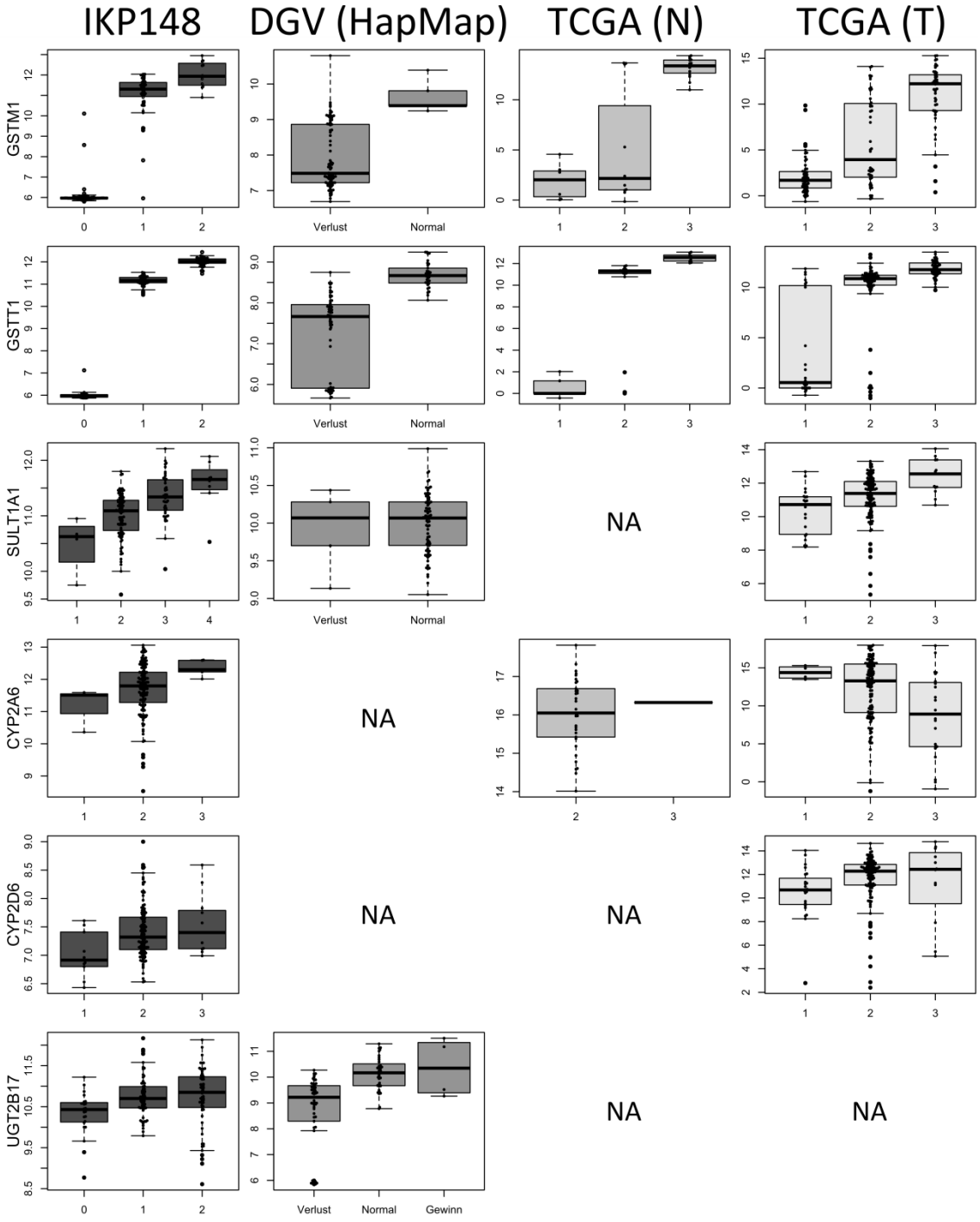


Abbildung 27: Boxplotanalysen zu CNVs, die eine signifikante Assoziation zur mRNA Expression in den IKP148-Leberproben aufwiesen, im Vergleich zu den anderen Kohorten. Von links nach rechts sind Expressionsdaten (Mikrochip) gegenüber CNV-Daten aus der NGS Analyse in den 150 Leberproben der Studie IKP148, Expression (Mikrochip, Stranger et al., 2007) und CNV-Daten (DGV) zu 90 HapMap-Proben (CEU= Einwohner mit Nord- und Westeuropäischer Vorfahren aus Utah, USA) und normalisierte RNA-Seq Expressionswerte gegenüber den extrahierten CNV-Segmenten aus Affymetrix Chip 6.0 Daten

von gesunden (N, n=34) und Tumor (T, n=153) Lebergewebeproben des TCGA-Projektes dargestellt. p-Werte zum Schaubild finden sich in Tabelle 10.

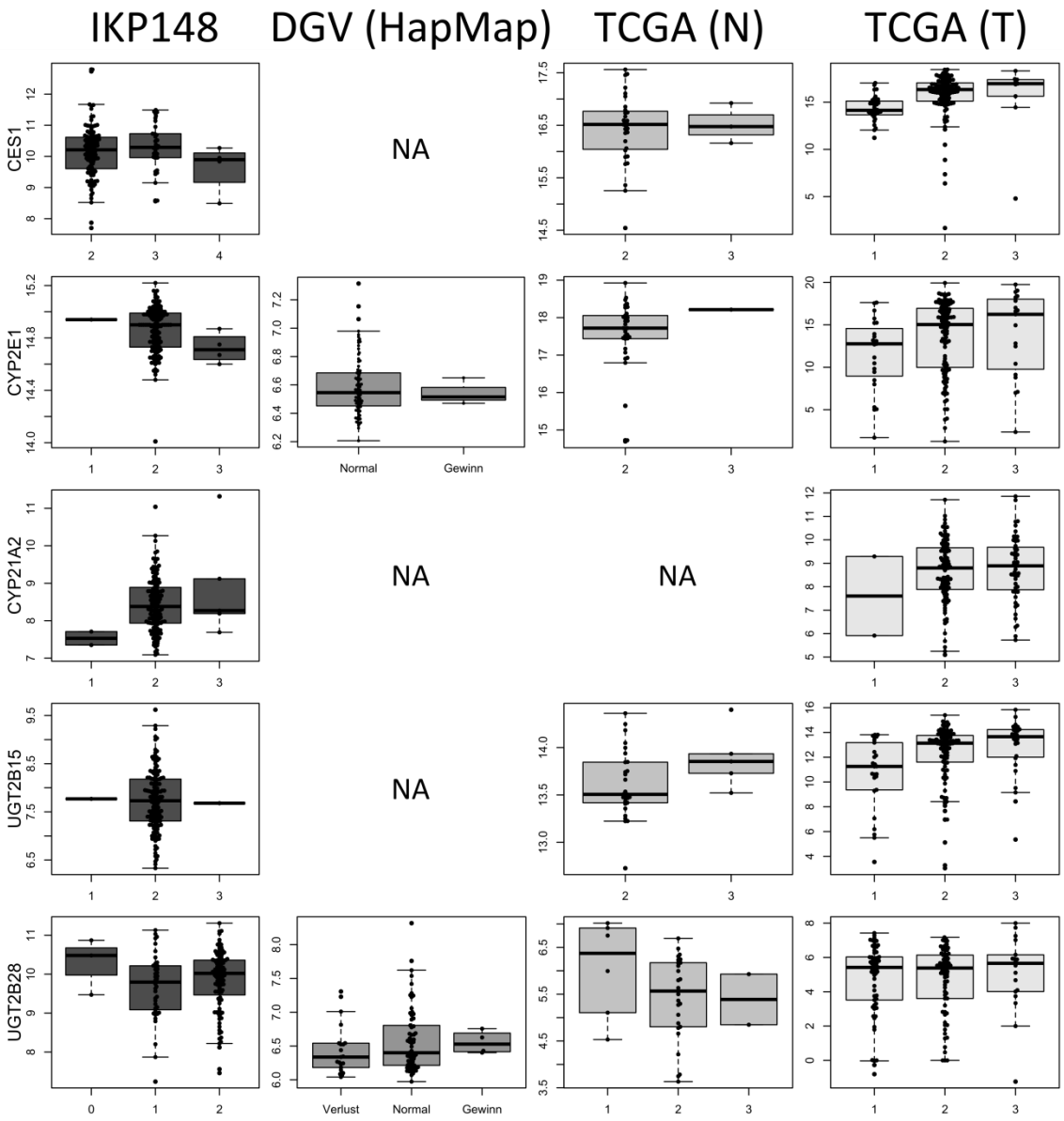


Abbildung 28: Boxplotanalysen zu CNVs, die keine Assoziation zur mRNA in den IKP148-Leberproben aufwiesen, im Vergleich zu den anderen Kohorten. Von links nach rechts sind Expressionsdaten (Mikrochip) gegenüber CNV-Daten aus der NGS Analyse in den 150 Leberproben der Studie IKP148, Expression (Mikrochip, Stranger et al., 2007) und CNV-Daten (DGV) zu 90 HapMap-Proben (CEU= Einwohner mit Nord- und Westeuropäischer Vorfahren aus Utah, USA) und normalisierte RNA-Seq Expressionswerte gegenüber den extrahierten CNV-Segmenten aus Affymetrix Chip 6.0 Daten von gesunden (N, n=34) und Tumor (T, n=153) Lebergewebeproben des TCGA-Projektes dargestellt. p-Werte zum Schaubild finden sich in Tabelle 10.

3.2.5 Funktionelle Analyse der *CYP2D6* und *CYP2A6* Genotypen in der IKP148-Leberkohorte

3.2.5.1 Funktionelle Analyse der *CYP2D6* Allele

Die durch die Feinkartierung ermittelten *CYP2D6* Allele in den IKP148-Leberproben wurden auf ihren Einfluss auf die Enzymaktivität (Propafenon 5'-Hydroxylierung, Gomes et al. 2009) untersucht. Die Donoren wurden entsprechend ihrer Anzahl an funktionalen *CYP2D6* Genkopien eingeteilt. Dabei wurden nicht-funktionelle Allele mit einer 0, Allele mit reduzierter Aktivität mit einer 0,5 und normal funktionierende Allele mit einer 1 bewertet und für jede Probe die Summe gebildet (Abbildung 29), wobei die Information zur Funktionalität aus der CYPalleles-Datenbank abgeleitet wurde. Eine univariate lineare Regressionsanalyse bestätigte die signifikante Assoziation ($p< 2E-16$) und zeigte, dass diese Einteilung der Genotypen mehr als die Hälfte der Aktivitätsvariabilität erklärte (adjustierter R^2=52%). Da Proben mit 0 und 0,5 oder 1 und 1,5 funktionellen Genkopien ein ähnliches Aktivitätsmuster aufwiesen, wurden die Leberspender entsprechend ihrer Metabolisierungseffizienz in langsame (die sogenannten *poor metabolizer*; PM), intermediäre (IM), normale (*extensive metabolizer*; EM) und schnelle (*ultra rapid*; UM) Metabolisierer eingeteilt. Die Mehrzahl der Spender wies einen IM Phänotyp (51%), gefolgt von einem EM Phänotyp (37%) auf. Seltener waren der PM (9%) und UM Phänotyp (3%).

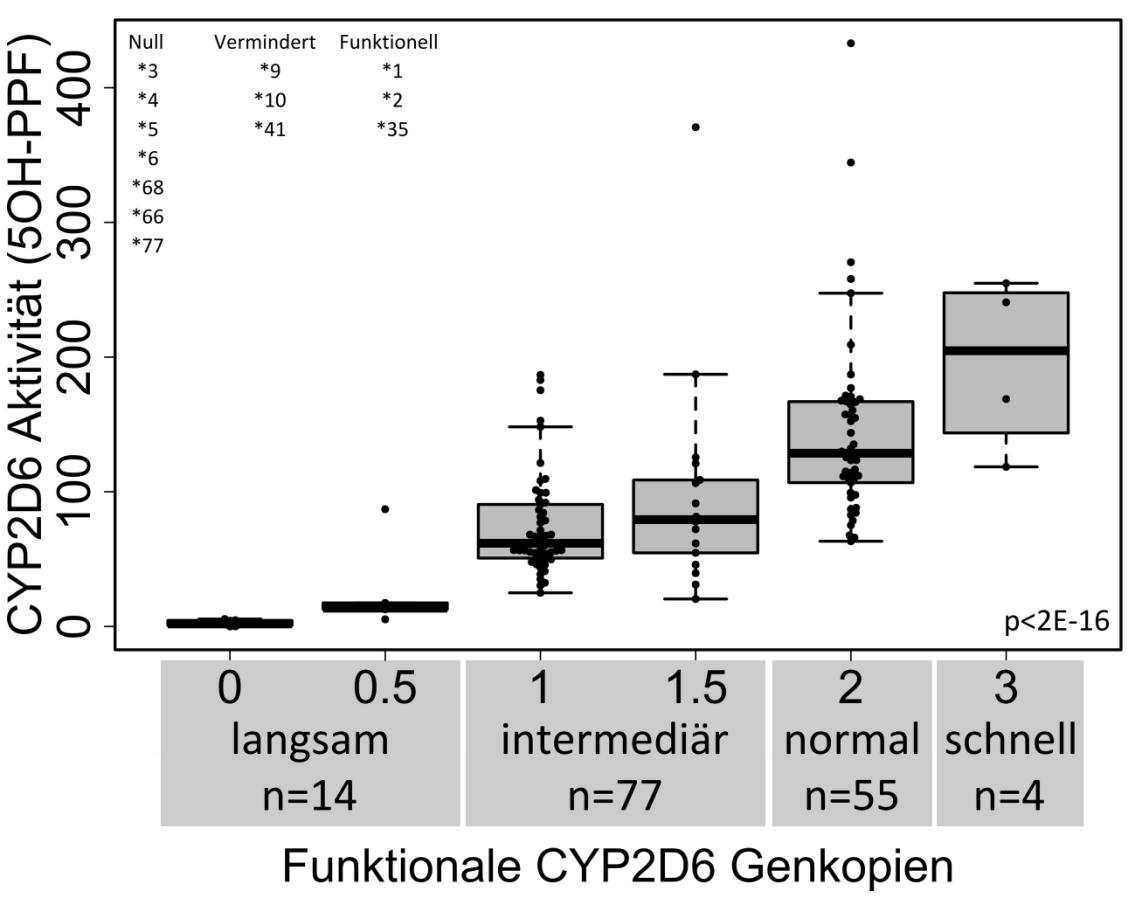


Abbildung 29: Assoziation der *CYP2D6* Genotypen mit der Enzymaktivität von CYP2D6. Die 150 Leberspender wurden entsprechend ihrer Anzahl an funktionellen Genkopien auf beiden Chromatiden gruppiert. Dabei entsprach ein nicht-funktionelles Allel einer 0, ein Allel mit verminderter Aktivität entsprach einer 0,5 und ein funktionelles Allel einer 1. Die Information dazu wurde von der CYP-Alle-Nomenklatur-Datenbank abgeleitet. Die mikrosomale Aktivität wurde mit dem Substrat Propafenon und per Massenspektroskopie bestimmt (Gomes et al. 2009). Die Assoziation wurde mit einer linearen Regression, in der die Verteilung der abhängigen Variable CYP2D6 Enzymaktivität mit einer Box-Cox Transformation einer Normalverteilung angepasst wurde (λ≈ 0,5), berechnet.

3.2.5.2 Funktionelle Analyse der bestimmten *CYP2A6* Allele

Der Analyse im vorigen Abschnitt folgend wurden auch *CYP2A6* Proben anhand ihrer funktionellen Genkopien gruppiert und der Einfluss auf die Enzymaktivität analysiert. In der Leberbank kamen keine Proben vor, die weniger als eine funktionelle Genkopie trugen. Obwohl kein medianer Unterschied in den Aktivitätswerten zwischen Proben mit 1 und 1,5 Kopien festzustellen war, konnte eine signifikante positive Genotyp-Phänotyp Beziehung beobachtet werden (p=2,14E-05; Abbildung 30). Der Genotyp erklärte 11% (adjustiert) der gesamten Variabilität. Nicht-genetische Faktoren hatten keinen bedeutenden Einfluss auf die CYP2A6 Aktivität.

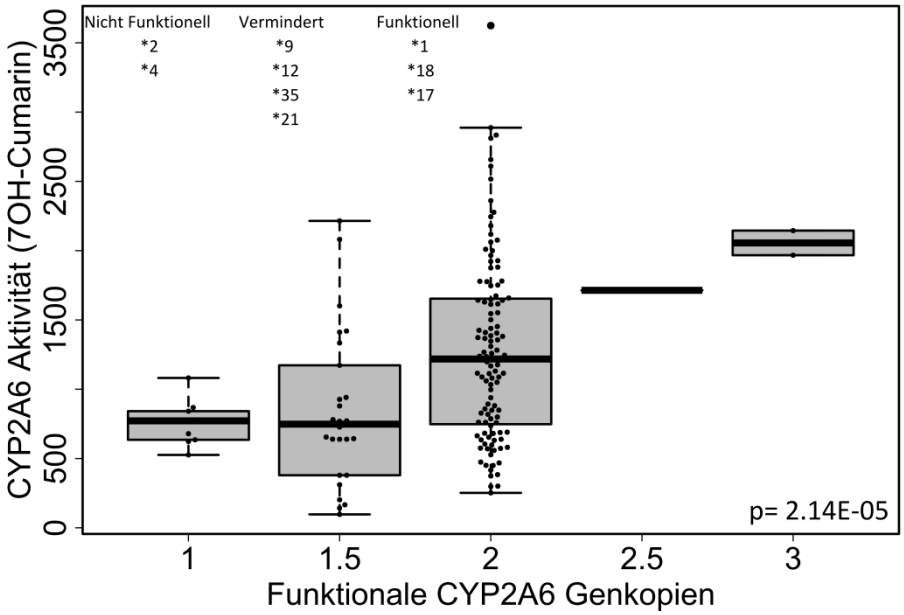


Abbildung 30: Assoziation der *CYP2A6* Genotypen mit der Enzymaktivität von CYP2A6. Die 150 Leberspender wurden entsprechend ihrer Anzahl an funktionellen Genkopien auf beiden Chromatiden gruppiert. Dabei entsprach ein nicht-funktionelles Allel einer 0, ein Allel mit verminderter Aktivität entsprach einer 0,5 und ein funktionelles Allel einer 1. Die Information dazu wurde von der CYP-Allel-Nomenklatur-Datenbank abgeleitet. Die mikrosomale Aktivität wurde mit dem Substrat Cumarin und per Massenspektroskopie bestimmt (Gomes et al. 2009). Die Assoziation wurde mit einer linearen Regression, in der die Verteilung der abhängigen Variable CYP2A6 Enzymaktivität mit einer Box-Cox Transformation einer Normalverteilung angepasst wurde ($\lambda \approx 0{,}4$), berechnet.

3.3 Funktionelle Untersuchung von genetischen Variationen im *CYP2E1* Gen

Im ersten Teil dieses Abschnittes wurde die oben beschriebene fehlende Assoziation zwischen *CYP2E1* CNVs und der mRNA Expression in allen drei Datensätzen detaillierter untersucht. Im zweiten Teil wurde ein SNP in der *CYP2E1* Genregion analysiert, der in einer Fall-Kontroll-Studie eines Kooperationspartners als Risikoallel für Schilddrüsenkrebs identifiziert worden war. Für beide Fragestellungen wurde die CYP2E1 mRNA Expression mit einem selektiven TaqMan Assay und qPCR erneut gemessen. Zudem wurde in den Leberproben die mikrosomale Proteinexpression mit einer WesternBlot Analyse und die Enzymaktivität mit dem Substrat Chlorzoxazon massenspektroskopisch quantifiziert.

Die in diesem Kapitel beschriebenen Ergebnisse zur CNV-Auswertung wurden in *The Pharmacogenetics Journal* (Tremmel et al., 2015), die SNP-Genotyp-Phänotyp Analyse in *Archives of Toxicology* als Originalarbeit publiziert (Pellé et al., 2016).

3.3.1 Deskriptive Analyse der CYP2E1 Phänotypen

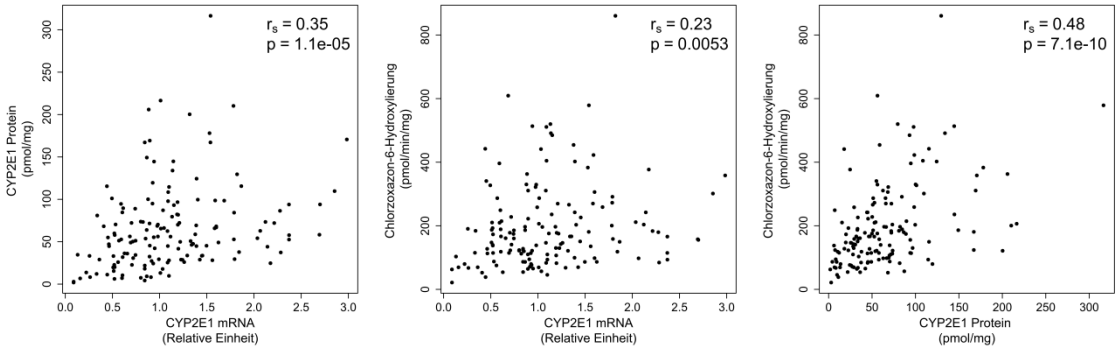


Abbildung 31: Untersuchung der Korrelation von hepatischen CYP2E1 Phänotypen in 150 humanen Leberproben. Die mRNA wurde mit qPCR und einem spezifischem Assay bestimmt. Microsomale Proteinexpression wurde mit einer WesternBlot Analyse und die CYP2E1 Enzymaktivität massenspektroskopisch quantifiziert. Der Spearmans Rangkorrelationskoeffizient und der p-Wert sind angegeben.

In den 150 Leberproben (IKP148) variierte die CYP2E1 mRNA Expression 33-fach, die Proteinwerte 168-fach und die Enzymaktivität 40-fach (Tabelle 11). Auch der Variationskoeffizient, der als normalisiertes Maß der Variabilität betrachtet werden kann, war für die Proteindaten am höchsten (76,6%). Alle Phänotypen waren nicht normalverteilt und korrelierten schwach miteinander, wobei die Proteinwerte mit den Enzym Aktivitäten am stärksten korrelierten (r_s=0,48) und eine schwache Korrelation zwischen mRNA und Aktivität zu beobachten war (p=0,23; Abbildung 31).

Tabelle 11: Variabilität der hepatischen CYP2E1 Phänotypen (n=150)

	mRNA Relative Einheit	**Protein [pmol/mg]**	**Chlorzoxazon-6-Hydroxylierung [pmol/min/mg]**
Minimum	0.09	1.88	21.51
Median	1.00	54.56	163.4
Maximum	2.98	316.5	860.4
Variabilität	33.1	168.2	40.0
Variationskoeffizient [%]	53.6	76.6	67.7

Mittels multivariater Analyse wurde der Zusammenhang der nicht-genetischen Faktoren und der CYP2E1 Phänotypen untersucht. Es ergab sich ein signifikanter Einfluss des Geschlechts und des Entzündungsbiomarkers CRP: Der Proteinwert war in Frauen signifikant erhöht (1,3-fach; p<0.001) und Leberspender mit erhöhtem CRP Spiegel

wiesen einen 3,2-fach erhöhten Proteingehalt (p<0.008) auf. Das Trinkverhalten von Alkohol zeigte ebenso wie alle weiteren untersuchten Faktoren keinen Einfluss auf die CYP2E1 Expression und Aktivität (Tabelle 12).

Tabelle 12: Einfluss der nicht-genetischen Faktoren auf die CYP2E1 Phänotypen

Nicht-genetischer Faktor	mRNA[1]	Protein[1]	Aktivität[1]
Geschlecht (M vs. W)	0,33	**0,001 ↑**	0,09
Alter (Aufsteigend)	0,82	0,23	0,84
Nikotinkonsum (Nein vs. Ja)	0,55	0,71	0,67
Alkoholkonsum (Nein vs. ≥1 x pro Woche)	0,08	0,08	0,63
Diagnose[2] (PLTvs. M)	0,62	0,54	0,60
Cholestasis (Nein vs. Ja)	0,26	0,87	0,39
GGT[3] (Normal vs. Erhöht)	0,39	0,65	0,66
CRP[4] (Normal vs. Erhöht)	0,11	**0,008 ↑**	0,91
Medikation[5] (Keine vs. Aktivator vs. andere)	0,50	0,96	0,22

1 p-Werte (ANOVA); Pfeile geben die Richtung an. p-Werte nach Benjamini Hochberg korrektur sind fett gezeichnet; **2** PLT, primärer Lebertumor; M, Metastase; **3** Gamma-glutamyl transferase; **4** C-reactive protein; **5** Medikation vor der Operation, Einteilung als P450 Aktivatoren nach (Klein et al., 2010).

3.3.2 Korrelationsanalyse der *CYP2E1* CNVs mit mRNA, Protein und Aktivitätsdaten von CYP2E1

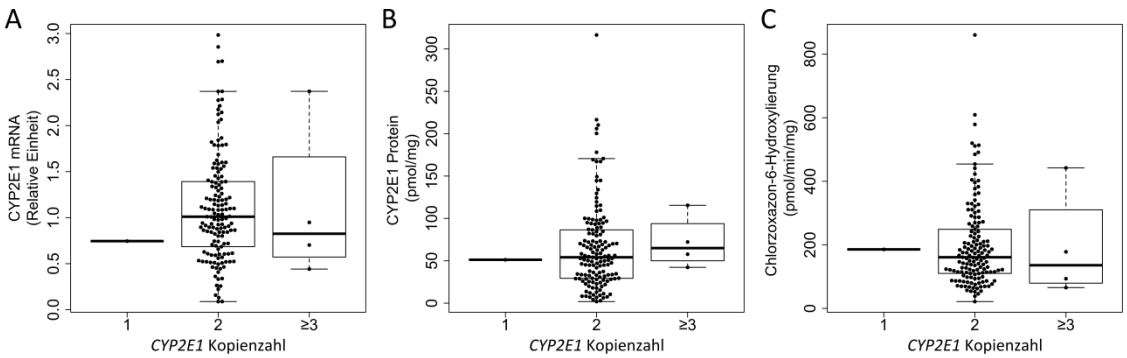


Abbildung 32: CYP2E1 Genotyp-Phänotyp Analyse. Zusammenhang von **A** mRNA, **B** Proteinexpression und **C** Enzymaktivität und der Kopienzahl von *CYP2E1*.

CYP2E1 mRNA-, Proteinexpression und Enzym Aktivitätsdaten wurden entsprechend der *CYP2E1* Kopienzahl gruppiert und analysiert. Wie in Abbildung 32 zu erkennen ist, zeigten Proben mit abweichender Kopienzahl im Vergleich zu Proben mit normaler Kopienzahl von zwei weder eine verminderte noch eine signifikant erhöhte mRNA und

Protein Expression oder Aktivität. In Trägern einer Duplikation war die mediane mRNA Expression 0,84-fach erniedrigt. Der mediane Proteinwert war 1,2-fach erhöht, wohingegen die Aktivitätsdaten 0,8-fach erniedrigt waren. Alle Veränderungen waren mit einem jeweiligen p-Wert über 0,4 nicht signifikant. Der Leberspender, der eine *CYP2E1* Deletion trug, hatte ebenfalls keine abweichende mRNA und Protein Expression oder Aktivität. Eine multivariate Analyse, bei der die nicht-genetischen Faktoren eingeschlossen wurden, bestätigte den fehlenden signifikanten Einfluss der *CYP2E1* CNVs auf die untersuchten Phänotypen (Ergebnisse nicht gezeigt).

3.3.3 CNV-gekoppelte SNPs als mögliche Faktoren eines Dosis-Kompensationsmechanismus

Das zunächst überraschende Ergebnis des fehlenden Zusammenhangs zwischen Kopienzahl und Phänotyp konnte zunächst nicht erklärt werden. Wie in der Einleitung bereits angedeutet, kommen verschiedene Kompensationsmechanismen in Frage. Eine Untersuchung von funktionellen SNPs, die zur Duplikation gelinkt sind, ließ sich in HapMap-Proben leicht bewerkstelligen.

Die LD-Analyse zwischen SNPs in der Region 10q26.3 und der *CYP2E1* Duplikation in HapMap-Proben mit europäischem Ursprung identifizierte 35 SNPs die signifikant ($r^2>0.6$) zur Duplikation gelinkt waren. Die SNPs waren auch untereinander zu 100% gelinkt und mit einer jeweiligen Allelfrequenz unter 5% selten (Abbildung 34A).

Mittels einer *in silico* Vorhersage wurde versucht, molekulare Mechanismen des Einflusses der SNPs in der *CYP2E1* Genregion vorherzusagen. Dazu wurden die Online-Tools Haploreg (Ward and Kellis, 2011) und der UCSC Genombrowser (Kent et al., 2002), welche beide auf Daten des *Encyclopedia of DNA Elements* (ENCODE) Projektes (Consortium, 2012b) zugreifen, benutzt. Zwei aufwärts von CYP2E1 gelegenen SNPs (rs192888328, rs6413423) und drei intronischen SNPs (rs72862138; rs41299414; rs28371747) lagen in DNase I hypersensitiven Regionen, CpG Inseln oder Transkriptionsfaktorbindestellen. Unter anderem waren die Bindestellen der Transkriptionsfaktoren PPARα, ein bekannter Kernrezeptor des Arzneimittelmetabolismus und USF1 (*Upstream Transcription Factor 1*) betroffen. Außerdem lieferten ChIP-Seq Daten den Hinweis, dass rs41299414 innerhalb einer RNA-Polymerase II Bindestelle lag und diese negativ beeinflusst. Die Intronregion der zwei SNPs (rs72862138; rs41299414) wies zusätzlich eine Histonmodifizierung auf, die als Indiz für einen schwachen Promoter gilt (Abbildung 34B). Der SNP innerhalb der 3'UTR (rs7081484; C>T) veränderte laut vorhersage Bindestellen für die miRNAs miR-4450, miR-4525,

miR-5010-5p und miR-506-5p. Insgesamt konnte daher auf einen SNP-bedingten negativen Einfluss auf die Expression geschlossen werden.

Um diese Hypothese zu überprüfen, wurde ein Taq-SNP (rs7081484, C>T, 3'UTR) ausgewählt und in den Leberspendern mit Hilfe einer TaqMan Sonde genotypisiert. Die Genotypen befanden sich im Hardy-Weinberg Gleichgewicht und waren wie in Individuen mit Europäischer Abstammung des 1000 Genomprojektes selten (f= 0,02). Alle Träger der *CYP2E1* Duplikation trugen das alternative T-Allel. Darüber hinaus kamen zwei Spender vor, die nur das heterozygote T-Allel, nicht aber die *CYP2E1* Duplikation trugen. In diesen zwei Spendern war die mRNA (p= 0,02), die Protein Expression (p=0,09) und die Aktivität (p= 0,4) im Vergleich zu Proben ohne Variante und CNV erniedrigt (Wilcoxon-Vorzeichen-Rang-Test; Abbildung 33). Anzumerken ist, dass in der LD-Analyse in Afrikanern kein zur Duplikation gelinkter SNP gefunden wurde. In Asiaten war auf Grund einer zu geringen Probenanzahl mit Duplikation keine Analyse möglich.

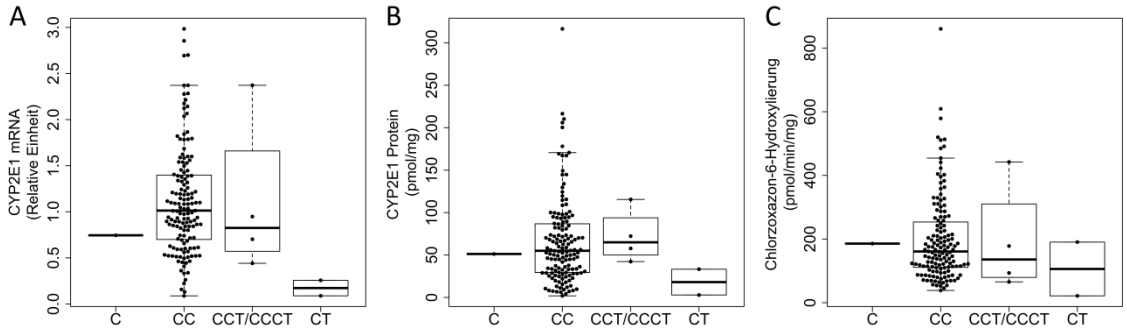


Abbildung 33: CYP2E1 Genotyp-Phänotyp Analyse. Zusammenhang von **A** mRNA, **B** Proteinexpression und **C** Enzymaktivität und der Anzahl an SNP-Allelen (rs7081484, C>T, 3'UTR), die durch die Kombination von SNP- und CNV-Genotypen bestimmt wurde.

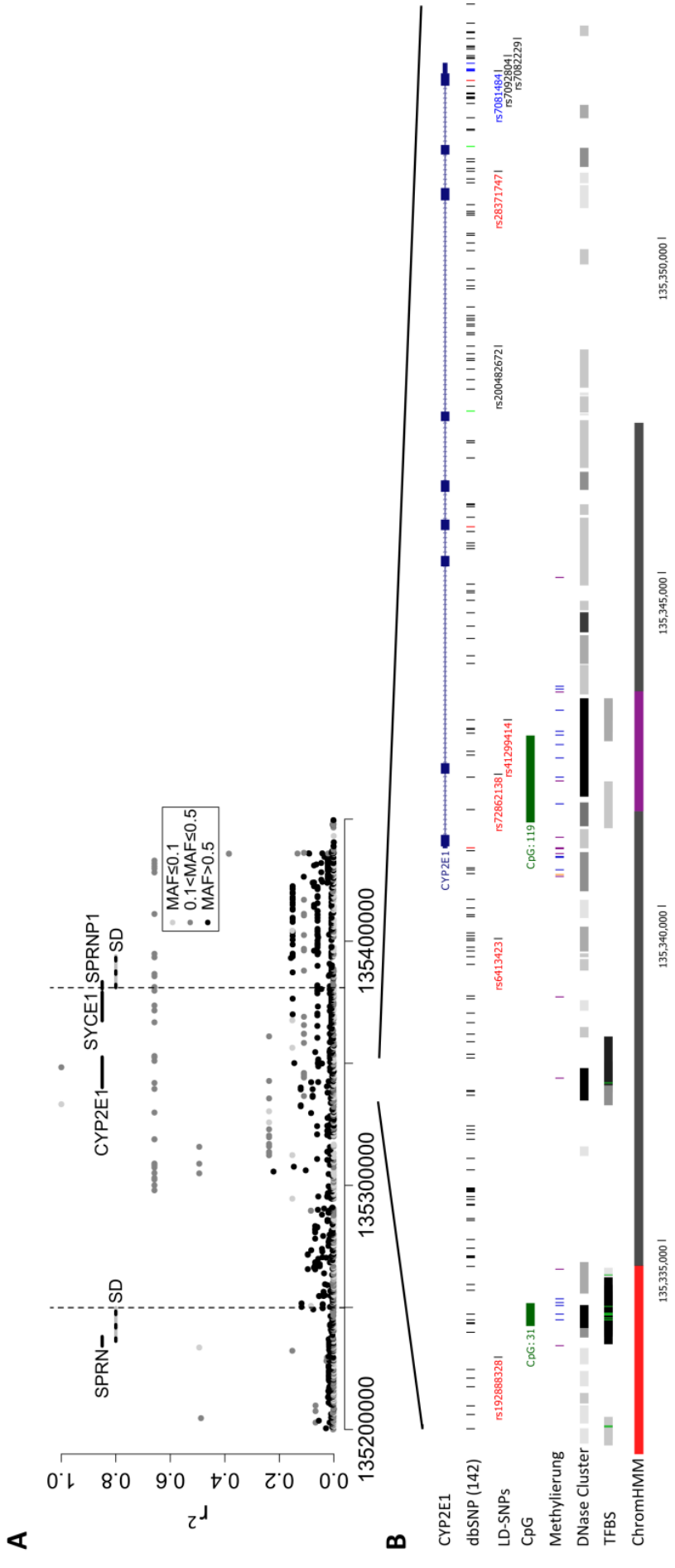

A
B
SPRN
SD
CYP2E1
SYCE1
SPRNP1
SD
MAF≤0.1
0.1<MAF≤0.5
MAF>0.5
r²
0.0
0.2
0.4
0.6
0.8
1.0
135200000
135300000
135400000
CYP2E1
dbSNP (142)
LD-SNPs
CpG
Methylierung
DNase Cluster
TFBS
ChromHMM
rs1928888328
CpG:31
rs64113423
rs72862136
rs412999414
CpG:119
rs200482672
rs283717471
rs70814864
rs70928041
rs70822291
135,335,000
135,340,000
135,345,000
135,350,000

Abbildung 34: LD-Diagramm und funktionelle *in silico* Analyse der SNPs in der *CYP2E1* Genregion. **A** LD zwischen der *CYP2E1* Duplikation und den SNPs in der Rgion 10q26.3. r^2-Werte wurden mit Daten des 1000G Projektes in Proben europäischer Herkunft mit der Software PLINK berechnet. Die vermeintlichen Bruchpunkte der CNVs nahe der segmentalen Duplikationen sind als horizontal gestrichelte Linien angedeutet. **B** Ausschnitt der *CYP2E1* Region. Von oben nach unten dargestellt sind alle in **dbSNP** vorkommende SNPs (Version 142); **LD-SNPs**, die neun SNPs, die zur Duplikation gelinkt sind und rot markiert wenn sie mit Daten des ENCODE Projektes zusammenfallen oder in blau wenn sie in miRNA Bindestellen liegen. **CpG**, CpG Inseln. **Methylierung**, Status der Methylierung der CpG Dinukleotide. DNase-I-Hypersensitive Bereiche (**DNase Cluster**); Transkriptionsfaktorbindestellen (**TFBS**) und eine Zusammenfassung der Chromatineigenschaften und Histonmodifikation (**ChromHmm**). In Rot ist ein schwacher Promoter, in Grau eine durch Polycomb reprimierte Region und in Violett ein inaktiver bis schwacher Promoter dargestellt. Die Übersicht wurde mit dem UCSC Genombrowser erstellt (GRCh 37; hg19).

3.3.4 Assoziation zwischen *CYP2E1* Polymorphismen und dem Schilddrüsenkrebsrisiko

Differenzierter Schilddrüsenkrebs (DTC) entsteht durch ein komplexes Zusammenspiel zwischen genetischen und Umweltfaktoren wie ionisierender Strahlung, untypischen Hormonspiegeln und vorangegangenen Schilddrüsenerkrankungen. Kürzlich wurde in Nagern *in vivo* gezeigt, dass über die Nahrung zugenommenes Acrylamid (AA) DTC auslösen kann (Beland et al., 2013). Weil Acrylamid in der Leber hauptsächlich von CYP2E1 zum aktiven Glycidamid, welches ebenfalls als genotoxische Substanz gilt, metabolisiert wird (Settels et al., 2008), wurde untersucht, ob *CYP2E1* Polymorphismen das Risiko für Schilddrüsenkrebs verändern. Dazu wurden vier tag-SNPs (rs9418982, rs743534 & rs7092584, rs2480258), die zu Polymorphismen im kompletten *CYP2E1* Genbereich gelinkt waren, in einer DTC Fall-Kontroll-Kohorte des Kooperationspartners genotypisiert (1117 DCT Patienten und 2779 Kontrollen; Informationen zur Kohorte in Pellé et al., (2016). Es stellte sich heraus, dass der tag-SNP rs2480258 (G>A), dessen Haplotyp weitere SNPs im hinteren Genbereich von Intron acht bis zur 3'UTR umfasste, signifikant zum Krebsrisiko assoziiert war. Das Risiko einer Erkrankung für das A Allel lag bei einem Quotenverhältnis (OR) von 1,17 (95% CI=1,04-1,31; p=0,008). Homozygote A Allel Träger wiesen dabei das höchste Risiko einer Erkrankung auf (OR=1,68; 1,20-2,35; p=0,002). Um den molekularen Mechanismus hinter diesem Ergebnis abzuschätzen wurde in 149 Leberproben (IKP148) eine Genotyp-Phänotyp Analyse mit den CYP2E1 Phänotypen mRNA und Protein Expression und Enzymaktivität durchgeführt. In den Leberproben waren Genotypen des SNP rs2515642, der signifikant zum Risiko-SNP gelinkt war (r2= 0.96), vorhanden. Die Genotypen, die mit einem SNP-Mikrochip bestimmt wurden (Schröder et al., 2013), waren signifikant mit reduzierter Expression und Aktivität assoziiert (Abbildung 35A). Aus imputierten SNP-Daten (siehe Methodenteil 2.2.4.7) konnte für rs2480258 der Genotyp

für alle Leberspender extrahiert werden. Die Frequenz der Genotypen (GG: 68%; GA: 29%; AA: 3%) unterschied sich nicht im Vergleich zur Frequenz in der DTC Kohorte. Im Vergleich zum homozygoten G Allel war das A Allel mit einer geringeren mRNA (GA: 0,83-fach, p=0,03; AA: 0,33-fach, p=0,003), Proteinexpression (GA: 0,74- fach, p=nicht signifikant; AA: 0,47- fach, p=0,03) und der Enzymaktivität (GA: 0,79- fach, p=0,03; AA: 0,8- fach, p= nicht signifikant; Abbildung 35B) assoziiert. Eine multivariate Analyse (lineares Regressionsmodell; mRNA, λ=0.45; Protein, λ=0.3; Aktivität, log-Transformation) mit den Variablen Geschlecht (weiblich vs. männlich), Alter, Nikotin- und Alkoholkonsum (ja vs. nein), eingenommene Medikamente (ja vs. nein), CRP (erhöht vs. normal), Cholestase (ja vs. nein) und der Diagnose (Primärtumor vs. Metastase), die zur Biopsie führte, bestätigte die signifikante Assoziation des A Allels mit niedriger mRNA (Additives Modell; p=0.01) und Proteinexpression (Additives Modell; p=0.01) und einer verminderten Chlorzoxazon 6-Hydroxylierung (Additives Modell; p=0.04).

Interessanterweise zeigte eine *in silico* Vorhersage mit Haploreg und PolymiRTs für die SNPs im LD-Block des tag-SNPs und Risko-Alleles rs2480258, dass zwei SNPs in der 3'UTR (rs24080257 und rs2480256) vorhergesagte Bindestellen für die miRNAs hsa-miR-5582-3p, has-miR-570-3p, hsa-miR-510-3p und hsa-miR-203a verändern.

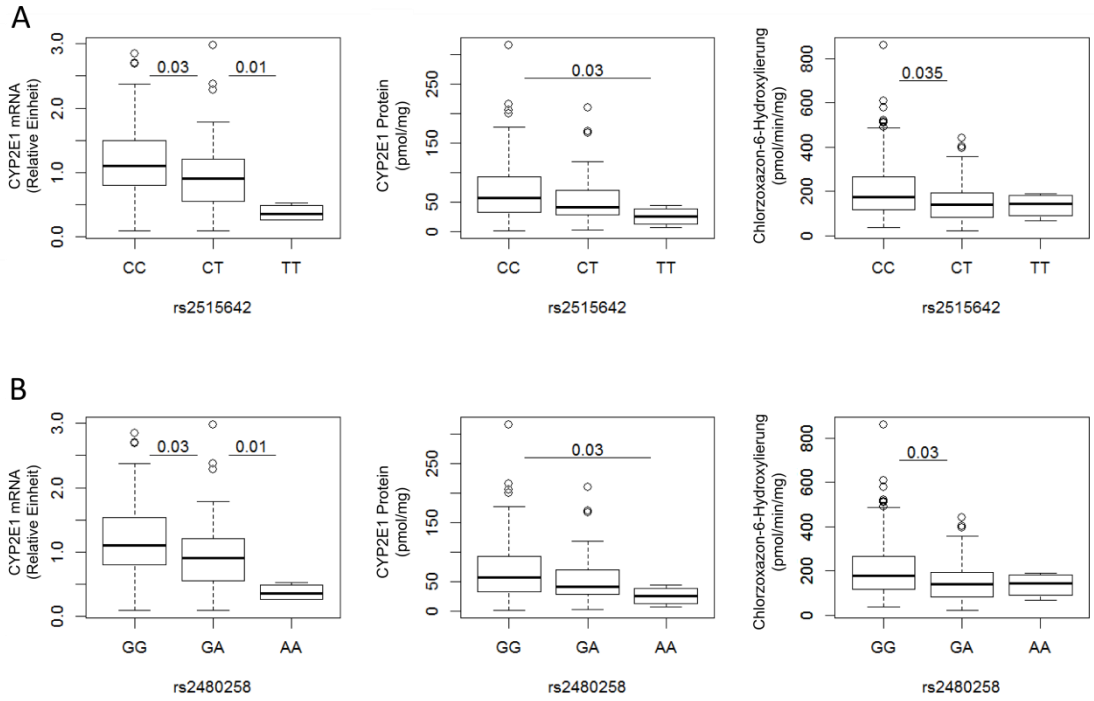


Abbildung 35: CYP2E1 Genotyp-Phänotyp Assoziationsanalyse in 149 humanen Leberspendern. Relative CYP2E1 mRNA, Proteinexpression und Enzymaktivität sind gegen Genotypen der SNPs **A** rs2515642 (Illumina Mikrochip) und **B** rs2480258 (imputiert) aufgetragen. Statistisch signifikante Unterschiede sind eingetragen (Wilcoxon-Vorzeichen-Rang-Test; p< 0,05).

3.4 Assoziation zwischen *SULT1A1* Genvarianten und Methyleugenol DNA-Addukten

Methyleugenol ist ein natürlicher Bestandteil von verschiedenen Gewürzpflanzen, Früchten und ätherischen Ölen (z.B. Fenchel, Estragon, Lorbeer, Gewürznelken und Basilikum). In Nagern wurde gezeigt, dass die Natursubstanz von Sult1a1 durch eine Sulfat-Konjugation aktiviert und dadurch genotoxisch wird und somit verschiedenste Krebsentitäten hervorrufen kann (Herrmann et al., 2012; National Toxicology Program, 2000). Im Detail bildet dabei das aktive Produkt 1-Sulfooxymethyleugenol ein DNA-Addukt, indem es überwiegend die Base Guanosin kovalent bindet (Herrmann et al., 2013). Eine Kooperationsarbeit lieferte erstmalig den Hinweis, dass auch in DNA von humanen Leberproben Methyleugenol DNA-Addukte zu finden sind (Herrmann et al., 2013). Deswegen wurde im Folgenden untersucht, ob die oben beschriebenen *SULT1A1* CNVs zur Variabilität des Methyleugenolmetabolismus beitragen und die DNA-Adduktmenge beeinflussen. Hierfür wurde in 121 Leberproben der Studie IKP148 zusätzlich zu den vorhandenen SULT1A1 mRNA Daten (Mikrochip Expressionsdaten; Sonden ID: ILMN_1656900) die Proteinexpression (WesternBlot Analyse in Leberzytosolfraktionen) bestimmt, sowie vom Kooperationspartner das Methyleugenol DNA-Adduktlevel per Massenspektroskopie gemessen. Die anschließende Assotiationsanalyse ist im Folgenden beschrieben.

3.4.1 Deskriptive Analyse der SULT1A1 mRNA und Protein Expression sowie der DNA-Adduktspiegel

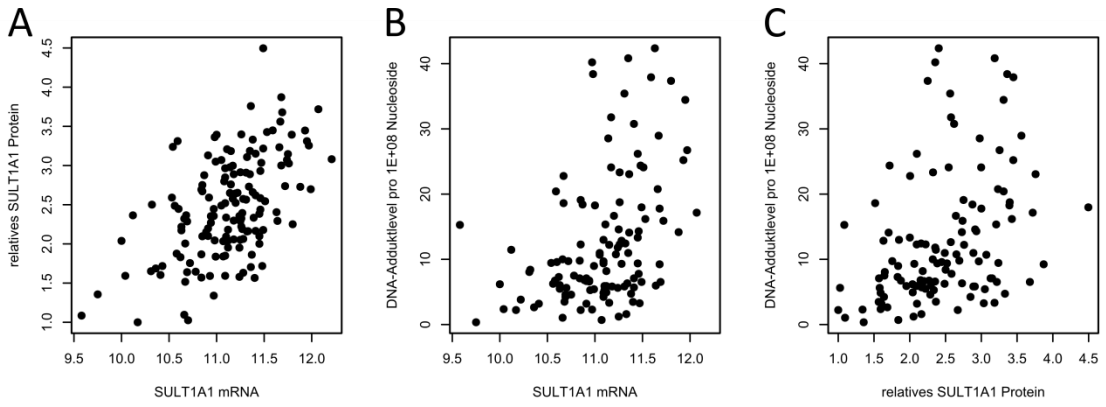


Abbildung 36: Korrelationsanalyse zwischen SULT1A1 mRNA (Mikrochip; Sonden ID ILMN_1656900), Proteinexpression (WesternBlot Analyse) und Methyleugenol DNA-Adduktwerten (Massenspektroskopische Auswertung) in 121 humanen Leberproben. Folgende Spearman's Rangkorrelationskoeffizienten und p-Werte wurden berechnet **A** r_s=0,5, p=1E-8; **B** r_s=0,43, p=1,1E-6; **C** r_s=0,44, p=6,6E-7.

Der Variationskoeffizient für mRNA und Protein betrug 30% und 27,6% wobei ein 5,6 bzw. ein 4,5- facher Unterschied zwischen höchstem und niedrigstem Expressionslevel zu beobachten war. Die DNA-Adduktmessungen variierten mit einem Koeffizienten von 79% und einem 122-fachem Unterschied zwischen Minimum und Maximum. Die SULT1A1 mRNA und Proteinwerte korrelierten signifikant miteinander (Abbildung 36; r_s=0,5). Nicht-genetische Faktoren wie die Altersangabe, das Geschlecht, die Diagnose vor der Operation, sowie der Spiegel des Leberfunktionsbiomarker CRP und der Alkohol und Nikotinkonsum hatten keinen Einfluss auf die SULT1A1 Expression (univariate nichtparametrische Tests). Auch die DNA-Adduktwerte waren kaum durch die nicht-genetischen Faktoren beeinflusst. So waren die Addukte in Patienten mit einem hohen Spiegel des Entzündungsbiomarker CRP leicht erniedrigt. Diese Änderung war aber nicht signifikant (p= 0,37, Mann-Whitney-Test). Außerdem nahmen die Adduktwerte mit dem Alter der Spender ab (r_s= -0,26, Spearman Test). Allerdings muss bemerkt werden, dass in der Leberkohorte Spender mit einem hohen Alter überrepräsentiert sind. Die positive Korrelation zwischen mRNA Expression (r_s=0,43) und Adduktmessungen sowie die Korrelation zwischen den Proteinwerten (r_s=0,44) und den DNA-Adduktmessungen lieferte den ersten Hinweis für die Bedeutung der SULT1A1 Enzymaktivität im Metyleugenolmetabolismus (Abbildung 36).

3.4.2 Beziehung zwischen den *SULT1A1* CNVs und der SULT1A1 mRNA und Proteinexpression

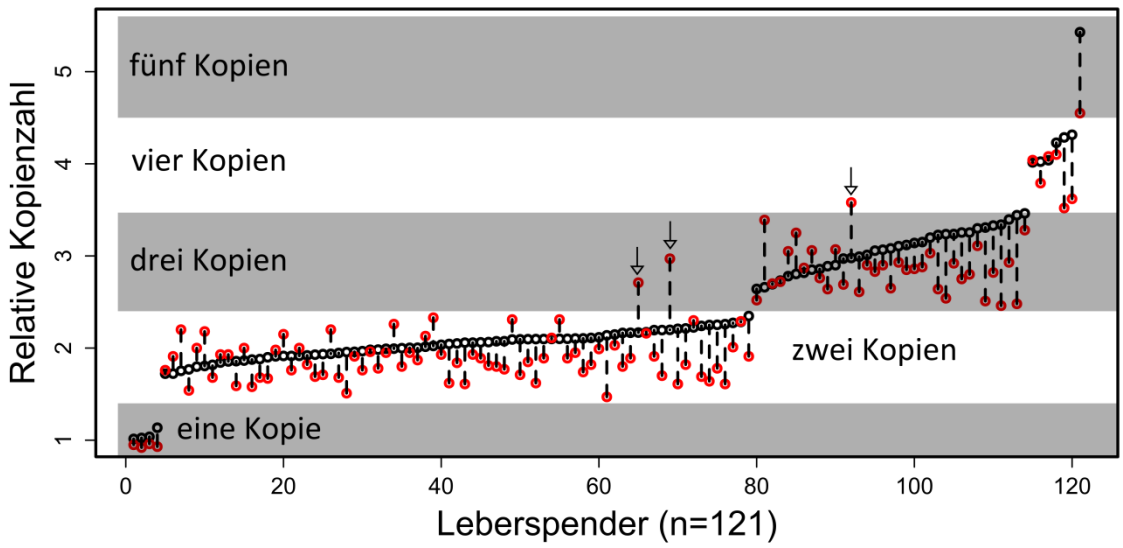


Abbildung 37: Ergebnis der *SULT1A1* Kopienzahlbestimmung in 121 Leberproben. Schwarze Kreise zeigen die jeweilige relative Kopienzahl der CNV-Detektion über die NGS-Daten (mit 2 multipliziert). Rote Kreise zeigen die relative Kopienzahl der qPCR-Methode mit der TaqMan CNV-Assay (Hs04461762_cn). Pfeile markieren Proben mit unterschiedlichen Ergebnissen zwischen den beiden Methoden.

Die Kopienzahlbestimmung der beiden eingesetzten Methoden war in drei der 121 Proben unterschiedlich (Abbildung 37). Da die NGS-Analysen auf Exon und Genebene in diesen drei Proben konsistente Ergebnisse lieferten, wurden in dieser Arbeit die Ergebnisse der NGS-Analyse für die folgenden Korrelationen verwendet. Die Spender wurden entsprechend ihrer *SULT1A1* Kopienzahl in vier Gruppen mit einer, zwei, drei und mehr als drei Genkopien eingeteilt und mit dem Wilcoxon-Vorzeichen-Rang Test die Assoziation zur SULT1A1 Expression berechnet. Wie in Abbildung 38C zu erkennen ist, lag eine signifikante Beziehung zwischen der Kopienzahl und der mRNA (Kruskal-Wallis Test; p=4,1E-7) vor. Im Vergleich zu Spendern mit zwei Kopien war der Median für mRNA Werte in Proben mit Deletion erniedrigt (nicht signifikant; p=0,05), wohingegen Spender mit drei Kopien signifikant erhöht mRNA (p=0,0002) exprimierten. Zusätzliche Kopien führten zu einem weiteren signifikanten Anstieg der mRNA-Expression im Vergleich zu Spendern mit drei Kopien (p=0,005). Eine multivariate Analyse, die die sechs nicht-genetischen Faktoren als Kovariable einschloss, ergab ebenfalls einen signifikanten Einfluss der SULT1A1 CNVs auf die SULT1A1 mRNA Expression (ANOVA und lineare Regression; p=5.3E-7).

Die Proteinisoformen der SULT1A Familie 1A1, 1A2 und 1A3 sind sehr homolog und haben eine ähnliche Molekülmasse von 34kDa, konnten aber mit einer WesternBlot Analyse elektrophoretisch aufgetrennt werden (Abbildung 38A). Durch das gleichzeitige Auftragen von rekombinant exprimierten Proteinproben konnte gezeigt werden, dass in den Leberproben kein SULT1A3 Protein vorkommt. Im Gegensatz dazu konnten die Isoformen SULT1A1 und SULT1A2 sowie dessen Variante SULT1A2*2 nachgewiesen werden. Dabei war SULT1A1 stärker exprimiert als beide SULT1A2 Proteinbanden (Abbildung 38A). Bereits der erste visuelle Eindruck der SULT1A1 Proteinbande auf dem Blot ließ eine Assoziation zur Kopienzahl vermuten (Abbildung 38B). Die Auswertung der relativen Proteinwerte, zeigte dann diesen starken und signifikanten Einfluss der SULT1A1 CNVs in einem univariaten Modell (Kruskal-Wallis Test; p=7,7E-11; Abbildung 38C) wie auch in einer multivariaten Regessionsanalyse (ANOVA und lineare Regression, p=8,6E-13). Proben mit einer Deletion exprimierten 0,71-fach weniger Protein (p=0,04). Eine heterozygote Duplikation führte zu drei Genkopien und einer 1,4-fach erhöhten Expression von SULT1A1 Protein (p=1E-9). Der Einfluss weiterer SULT1A1 Genkopien im Vergleich zu Proben mit drei Kopien war nicht mehr so prominent (1,1-fach) und nicht signifikant (Abbildung 38D).

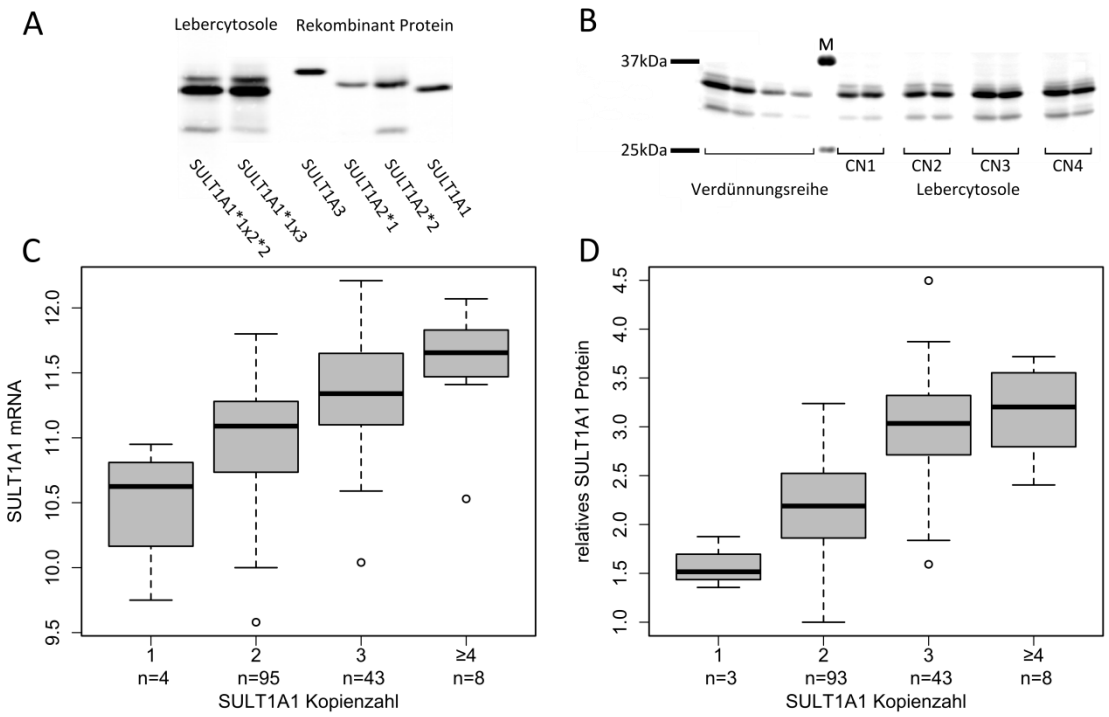


Abbildung 38: Korrelation zwischen *SULT1A1* CNVs und mRNA oder Proteinexpression. **A** Spezifitätsnachweis des eingesetzten Antikörpers. SULT1A3 wurde in den Leberzytosolen nicht detektiert. Die mittlere Bande wurde SULT1A1 zugeordnet. Die Banden ober- und unterhalb der SULT1A1 Bande migrierten auf gleicher Höhe wie rekombinantes SULT1A2 **B** Ein repräsentativer WesternBlot von 15 Leberzytosolproben mit unterschiedlicher *SULT1A1* Kopienzahl. Boxplot-Diagramme der vier *SULT1A1* Kopienzahlgruppen und **C** SULT1A1 mRNA oder **D** relativer SULT1A1 Proteinexpression. Für drei Leberproben konnte auf Grund von fehlendem Lebergewebe kein Proteinwert ermittelt werden.

3.4.3 Die Bedeutung der *SULT1A1* CNVs für den Phase II Metabolismus von Hydroxy-Methyleugenol

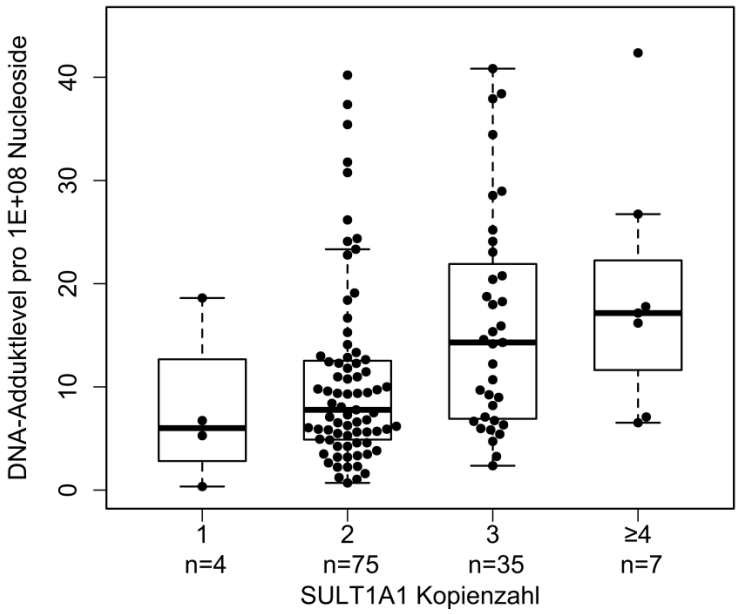


Abbildung 39: Assoziation zwischen der SULT1A1 Kopienzahl und den DNA-Adduktmessungen.

Wie in Abbildung 39 zu erkennen ist, erhöhte sich das DNA-Adduktlevel signifikant mit der Anzahl an *SULT1A1* Genkopien (Kruskal-Wallis-Test; p=0,005). Der stärkste Anstieg wurde zwischen Spendern mit zwei und drei Genkopien beobachtet (1,8-fach; p=0,003). Auch in einer multivariaten Analyse war die *SULT1A1* Kopienzahl signifikant zu den Adduktmessungen assoziiert (p=0,004). Der Einfluss auf die Variabilität von genetischem (CNV) und den nicht-genetischen Faktoren, die im univariaten Modell einen geringen Einfluss zeigten (Alter und CRP Spiegel), wurde mit einem linearen Modell erfasst. Der CNV-Einfluss betrug 12,6% (R^2 adjustiert =10%), was nach Alter (3,7%) und CRP Spiegel (1,2%) der größte Beitrag zur Variabilität war.

3.4.4 Einfluss des nicht-synonymen SNPs SULT1A1*2 (rs9282861, G638A Arg213His)

Neben den *SULT1A1* CNVs wurde gezeigt, dass auch das *SULT1A1*2* Allel (rs9282861, G638A, Arg213His) einen Einfluss auf die Proteinexpression nehmen kann (Raftogianis et al., 1999) und in Fall-Kontroll-Studien als Risikoallel gilt (Arslan et al., 2011; Rouprêt et al., 2007). Deswegen wurde im Folgenden der Einfluss des SNPs auf die SULT1A1 Expression und auf das DNA-Adduktlevel untersucht.

3.4.4.1 Genotyp- Phänotyp Analysen

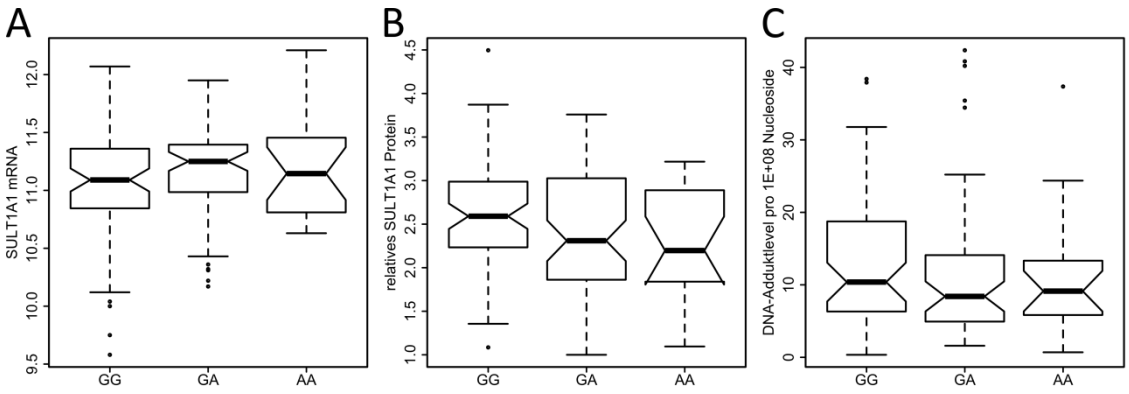


Abbildung 40: SULT1A1 mRNA, relative SULT1A1 Proteinexpression und Methyleugenol DNA-Adduktmessungen gruppiert nach dem Genotyp des nicht-synonymen SNP *SULT1A1*2* (rs9282861, G638A, Arg213His).

Die SNP Genotypen, die sich im Hardy-Weinberg-Gleichgewicht befanden, hatten einen signifikanten Einfluss auf die Proteinexpression (Kruskal-Wallis-Test; p=0,02). In Proben mit einem oder zwei A Allelen war die Proteinexpression 0,8-fach niedriger als in Proben mit homozygotem G Allel (Abbildung 40). Die Frage, welche genetische Determinante (*SULT1A1*2* Allel oder *SULT1A1* CNV) den stärkeren Einfluss auf die SULT1A1 Proteinexpression hat und ob noch weitere Varianten Einfluss nehmen, wurde mit einer kombinierten Analyse der CNV und SNP Genotypen geklärt.

In den Leberproben traten 12 verschiedene Allele auf, deren Vorkommen und Frequenz mit einer kürzlich erschienenen Studie vergleichbar waren (Tabelle 13).

Tabelle 13: Die Tabelle zeigt Häufigkeiten der Genotypen des nicht-synonymen SNPs (rs9282861; G>A) in Relation zur *SULT1A1* Kopienzahl in 121 humanen Leberproben (IKP148). Zum Vergleich sind Frequenzen einer weiteren Studie angegeben.

Kopienzahl	Genotyp	Diese Arbeit (n=121)	Hebbring et al., 2009 (n=761)
1	G	1,7%	2,7%
	A	1,7%	1,2%
2	GG	21,5%	22,3%
	GA	28,1%	31%
	AA	12,4%	11%
3	GGG	17,4%	12,7%
	GGA	9,9%	9,4%

Kopienzahl	Genotyp	Diese Arbeit (n=121)	Hebbring et al., 2009 (n=761)
	GAA	0,8%	-
	AAA	0,8%	0,4%
4	GGGG	3,3%	3,6%
	GGGA	1,7%	1,3%
5	GGGGG	0,8%	-

Die Gruppierung der Proben nach CNV-Status und SNP-Genotyp zeigte, dass der Einfluss des *SULT1A1*2* Genotyps auf SULT1A1 mRNA und Protein Expression oder den DNA-Adduktlevels nicht konsistent war (Abbildung 41). Die mRNA Werte waren in Proben mit zusätzlichem A-Allele tendenziell erhöht (nicht signifikant), wohingegen die Proteinexpression in Trägern von zwei *SULT1A1* Genkopien und heterozygoten SNP im Vergleich zu Trägern zweier normaler *SULT1A1*1* Genkopien signifikant erniedrigt war (Wilcoxon-Vorzeichen-Rang-Test; p=0,015). Die Adduktmessungen waren in keiner Gruppe signifikant verändert (Abbildung 41).

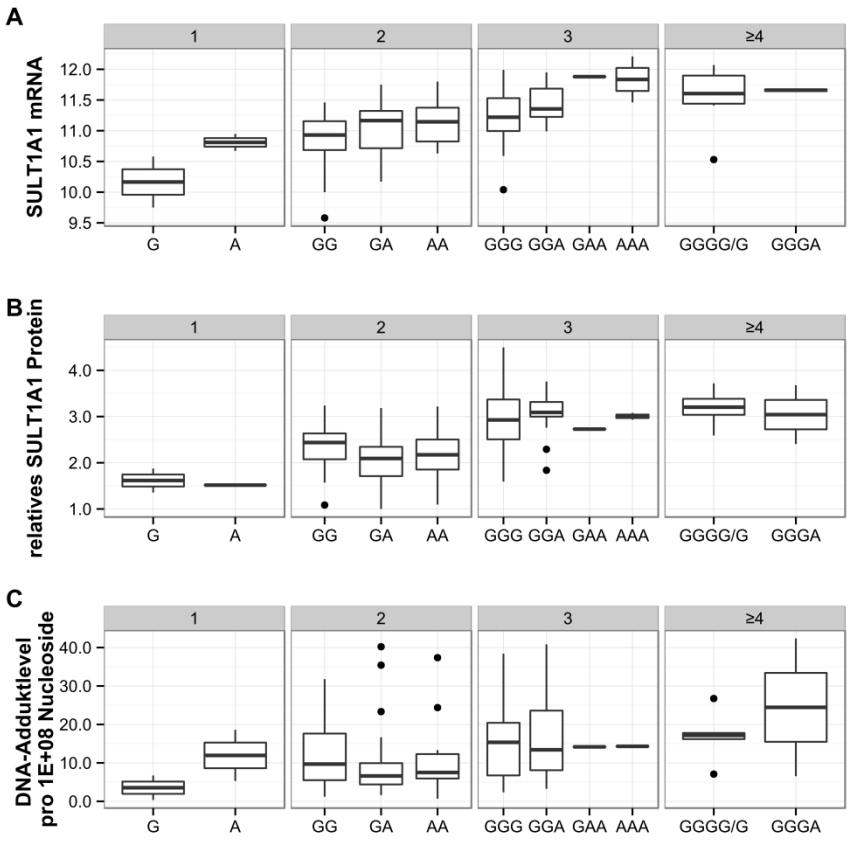


Abbildung 41: Einfluss des SNPs *SULT1A1*2* (rs9282861; C>T; Arg213His) in Kombination mit der *SULT1A1* Kopienzahl auf **A** die SULT1A1 mRNA, **B** die relative Proteinexpression und **C** den DNA-Adduktmessungen.

Zuletzt wurde mit Hilfe genomweiter SNP-Genotypen die Assoziation von *SULT1A1*2* und *SULT1A1* CNVs zur Proteinexpressionsvariabilität erneut berechnet und nach weiteren genetischen Einflüssen gesucht. Durch die Methode der Imputierung von SNP-Daten (siehe 2.2.4.7) standen final 7.538.834 SNP- und CNV-Genotypen zur Verfügung, für die jeweils mit einem additiven Modell die Assoziation zur Proteinexpression analysiert wurde. Das Ergebnis war eindeutig, wie im Manhattan-Diagramm der p-Werte zu erkennen ist (Abbildung 42). Nur die Assoziation der *SULT1A1* Kopienzahlvariation überschritt das genomweite Signifikanzniveau (p=5E-8) mit einem p-Wert von 6,2E-17. Der nicht-synonymen SNP rs9282861 war in dieser Analyse nicht signifikant zur Proteinexpression assoziiert (p=0,03). Es wurden keine weiteren genetischen Varianten, die zur Proteinexpression assoziiert waren, beobachtet.

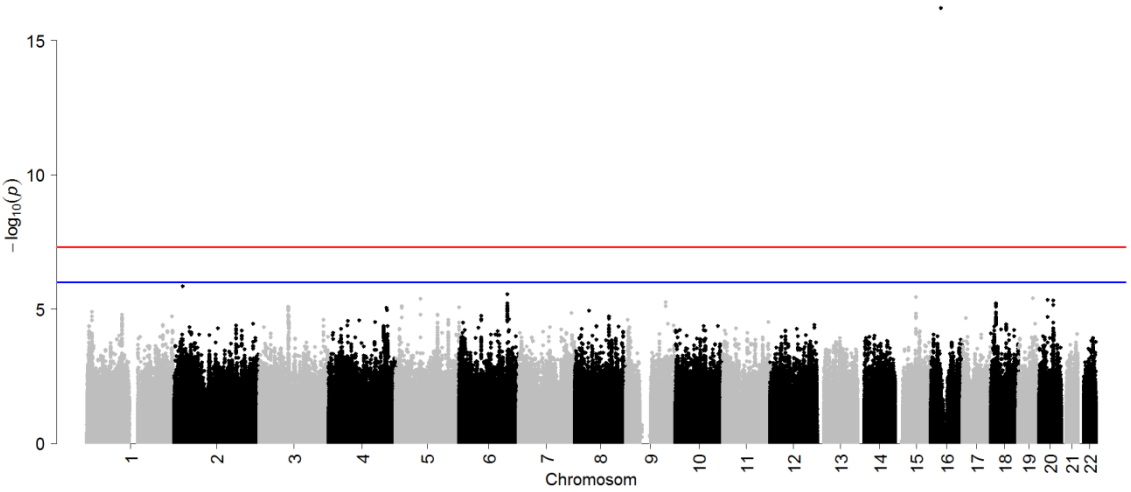


Abbildung 42: Manhattan-Diagramm der genomweiten Assoziationsanalyse für SULT1A1 Proteinexpression. Die blaue Linie zeigt eine suggestive Assoziationslinie (p= 1E-05), die rote Linie eine genomweite Signifikanz (p= 5E-08). Der einzelne Punkt im Bereich des Chromosoms 16 beschreibt die *SULT1A1* Kopienzahl, die als eins, zwei und mehr als zwei Kopien kodiert ist (p< 1E-15).

4 Diskussion

4.1 ADME-weite CNV-Bestimmung in humanen Leberspendern und HapMap-Proben

In dieser Arbeit wurde in drei unabhängigen Kohorten systematisch das CNV-Vorkommen in 340 ADME-Genen untersucht. Dabei wurde auf zwei online frei zugängliche Datensätze und auf eine Sammlung von150 humanen Lebergewebeproben mit zugehörigen Phänotypdaten (IKP148) zurückgegriffen. Im ersten Teil der Diskussion wird auf spezifische Fragen und allgemeine Ergebnisse zu den einzelnen Datensätzen und Methoden eingegangen. Die funktionelle Analyse der CNVs von ADME-Genen und detailliertere Analysen werden in den weiteren Teilen näher diskutiert.

4.1.1 Das CNV-Vorkommen von ADME-Genen in DGV

Zum Zeitpunkt dieser Untersuchung (Zugriff Oktober, 2014) waren in DGV-Daten zu CNVs hinterlegt, die durchschnittlich 70% des humanen Genoms und 92% der codierenden Bereiche abdecken. Diese beiden Werte waren weit höher als die von Redon und Kollegen (2006) beschriebenen 12% des Genoms und sollten kritisch hinterfragt werden. Zwar verspricht die Datenbank nur Varianten aufzunehmen, die in gesunden Proben gefunden wurden, aber wie oben bereits beschrieben, sind die Ergebnisse der CNV-Detektion sehr von der angewandten Methode und deren technischen Möglichkeiten abhängig. Ältere Verfahren wie z.B. BAC-, Oligo- oder SNP-Mikrochips der ersten Generationen neigen zu einer Detektion von zu langen Varianten und ungenauen Start- und Endpunkten (MacDonald et al., 2014).

Die beobachtete höhere Anzahl von gefundene Deletionen (13.912) im Vergleich zu den Duplikationen (2.793) in den ADME-Genloci aller in DGV hinterlegten Daten kann ebenfalls durch die Eigenheiten der einzelnen Methoden erklärt werden. Einerseits ist die Detektionssensitivität von Deletionen bei Oligo- und SNP-Mikrochip Verfahren höher als die Sensitivität für Duplikationen (Alkan et al., 2011). Und in der Tat wurde die Mehrzahl der gefunden Varianten in den ADME-Genregionen durch eines dieser beiden Verfahren entdeckt (Tabelle A1). Andererseits ist auch die CNV-Frequenz für die höhere Anzahl von detektierten Deletionen verantwortlich. In den Genregionen *GSTM1, GSTT1, UGT2B17* und *UGT2B28* kommen Deletionen weltweit mit einer Frequenz über 30% vor. Alle Methoden detektieren also in diesen Genbereichen mit höherer Wahrscheinlichkeit eine Deletion. Duplikationen mit ähnlichen Frequenzen über

30% sind in den ADME-Genen bisher nicht beschrieben und kommen auch in DGV nicht vor (MacDonald et al., 2014).

Es scheint, dass in manchen Genregionen leichter CNVs detektiert werden können als in anderen. Zum Beispiel fanden fast alle in dieser Arbeit berücksichtigten Studien in der *CYP2E1*, *GSTT1* und *UGT2B28* Genregion CNVs. Im Gegensatz dazu wurden etablierte CNVs von *CYP2D6* oder *SULT1A1* nur von wenigen Studien und mit zu niedrigen Frequenzen gefunden (Tabelle S1). Eine mögliche Erklärung ist die technische Limitation hinsichtlich der Anzahl und Position der Sonden im Genom verschiedener SNP-Mikrochips. Die Sondendichte des in den meisten Studien verwendeten SNP-Mikrochips (Affymetrix 6.0) ist zum Beispiel in der *CYP2D6* und *SULT1A1* im Vergleich zur CYP2E1 Region sehr niedrig und somit ist die Detektion in solchen Regionen unvollständig und fehleranfällig (Abbildung 43).

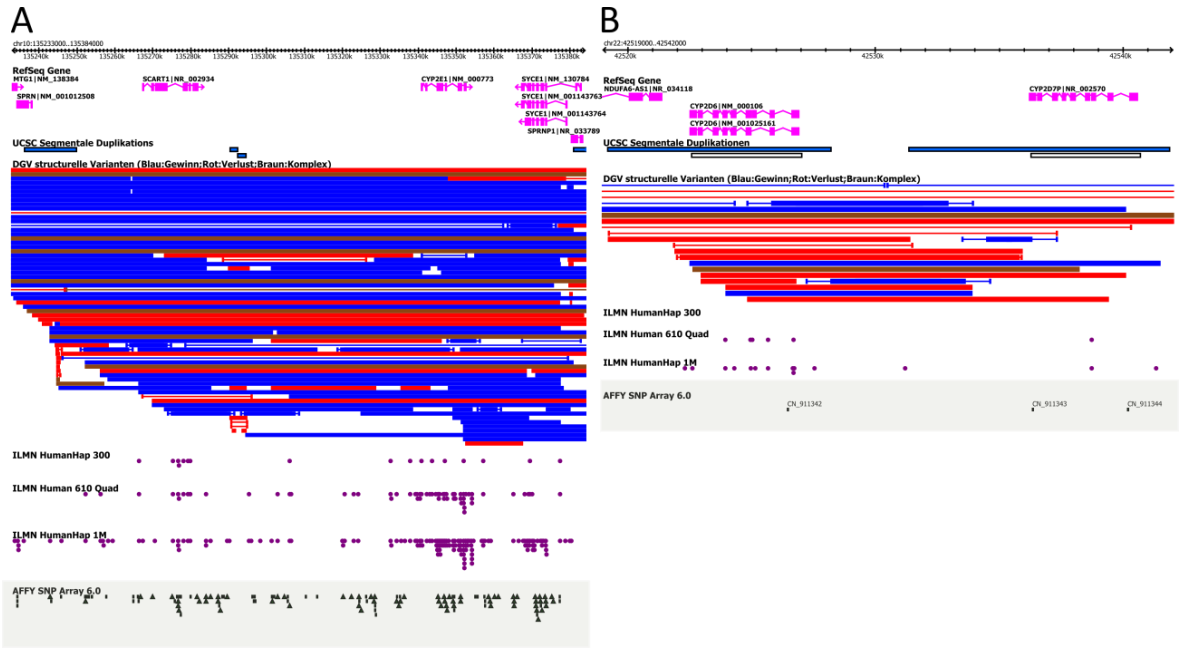


Abbildung 43: Übersicht der **A** *CYP2E1* und **B** *CYP2D6* Genregion und genomische Eigenschaften verschiedener SNP-Mikrochips. Von oben nach unten sind jeweils die genomische Position (GRCh37; hg19), die Genregionen (RefSeg), segmentale Duplikationen, die in DGV hinterlegten strukturellen Varianten und die Position der Sonden von vier verschiedenen Mikrochips für eine Genotypisierung gezeigt. Die Informationen zu den Sonden des Affymetrix 6.0 Mikrochips sind grau hinterlegt. Eine genauere Erklärung der weiteren Features findet sich in Abbildung 3.

Die populationsabhängige Auswertung der CNV-Verteilung in HapMap-Proben ergab ebenfalls mehr Deletionen als Duplikationen in den ADME-Genen. Allerdings war die CNV-Verteilung in den ADME-Gruppen unterschiedlicher als im kompletten DGV-Datensatz. Der größte Unterschied ergab sich für die Anzahl der gefundenen CNVs in Transportern (30% zu 11%). Es stellte sich heraus, dass zwei Studien, die keine

HapMap-Proben verwendeten, übermäßig viele CNVs in Transportern und ADME-Verwandten Genen fanden (siehe Tabelle S1; Cooper et al., 2011; Itsara et al., 2009). Ein möglicher Grund dafür könnte wieder das verwendete Verfahren zur CNV-Detektion oder die Zusammensetzung der Kohorte sein, die im ersten Fall nicht ausschließlich aus gesunden Probanden bestehen könnte. Die in den GST- und UGT2B-Genfamilien vertauschte Kopienzahl ist ein generelles Problem bei Verfahren, die nur eine Probe als Referenz verwenden und kann nur im Vergleich mit anderen Proben und Studien auffallen und behoben werden.

4.1.2 CNV-Bestimmung in TCGA-Leberproben

Die CNV-Bestimmung im TCGA-Datensatz beruht auf vom Konsortium prozessierten Daten eines SNP-Mikrochipexperiments (Affymetrix 6.0) von gepaarten Biopsien aus einem hepatozellulären Karzinom und aus gesundem Lebergewebe. Die in dieser Arbeit bestimmten CNVs in gesunden Proben überlappten zu 94% mit denen in DGV und einer Studie, die den gleichen Chiptyp verwendet hatte (Altshuler et al., 2010). Es wurden also größtenteils CNVs detektiert, die in DGV bereits beschrieben wurden. Dieses Ergebnis demonstriert aber auch die Verlässlichkeit des Verfahrens. Der Grad der Übereinstimmung lag in einem Bereich, den auch andere Studien als ausreichend betrachteten (McCarroll et al., 2008). Auch im TCGA-Datensatz wurde eine, durch die Methode bedingte (siehe Abschnitt 4.1.1), höhere Anzahl an Deletionen als Duplikationen gefunden. In Tumorproben wurde durch die signifikante Zunahme an langen Duplikationen dieser Unterschied nicht mehr beobachtet (Abbildung 13). Die generelle signifikante Zunahme an CNVs in Tumorzellen, besonders von Duplikationen in den Chromosomen 1, 5, 6, 8 und 20, wurde ebenfalls außerhalb TCGA beschrieben (Kan et al., 2013) und lässt sich durch eine erhöhte chromosomale Instabilität in Krebszellen erklären (Shibata and Aburatani, 2014). Es wurden CNVs in nahezu allen 340 ADME-Genen gefunden. Die am häufigsten auftretenden somatischen CNVs fanden sich in den Genen *NAT1* und *NAT2* (Deletionen) sowie *NR1I3* (CAR), *NR5A2*, *ESRRG*, *ARNT*, *CRPK* und *ALDH9A1* (Duplikationen) und waren zum Teil bereits mit TCGA-Daten aber auch in anderen HCC Kohorten und Studien beschrieben worden (Kim et al., 2015; Wang et al., 2013). Anzumerken ist, dass die Frequenzen der ermittelten CNVs im Vergleich zu den beiden Studien höher waren und CNVs von Genen des xenobiotischen Metabolismus bisher in HCC keine wichtige Rolle in der Krebsentstehung zugeschrieben wurden (Shibata and Aburatani, 2014). Da diese Ergebnisse zeigen, dass in HCC jedes ADME-Gen von strukturellen Varianten betroffen sein kann, würde es Sinn machen, in Zukunft in HCC Patienten diese Information in eine individualisierte Therapie einfließen zu lassen.

4.1.3 CNV-Bestimmung in der IKP148-Leberkohorte

Ein panelbasiertes ADME-Exonresequenzierungsprojekt (NGS) wurde zur CNV-Bestimmung in den 150 Leberproben der Studie IKP148 verwendet. Dazu wurde eine für diesen Datensatz angepasste Analyse der Abdeckung optimiert. Dieser Schritt war notwendig, weil die publizierten und häufig verwendeten Programme und Algorithmen entweder nur für Sequenzierungen des kompletten Genoms oder Exoms entwickelt wurden (Zhao et al., 2013). Andere NGS-Verfahren, die aus der Art der Anlagerung der gepaarten Sequenzierungen an das Referenzgenom CNVs vorhersagen (siehe Einleitung), waren aus diversen Gründen ungeeignet. So konnte nicht direkt auf die für diese Verfahren wichtigen Rohdaten zurückgegriffen werden. Zudem war die Rechnerleistung oder der im Haus verfügbare Festplattenspeicher nicht ausreichend. Ein weiterer Vorteil der weiter entwickelten Methode war die Quantifizierung der genauen bzw. relativen Kopienzahl, die mit den anderen Verfahren überhaupt nicht möglich ist. Durch das Erstellen einer Referenzprobe konnten falsch-positive Ergebnisse minimiert werden. Die Analyse auf Genebene, die alle Exons eines Gens zu einem medianen Wert zusammenfasste, wurde gewählt, weil sie mit hoher Wahrscheinlichkeit nur eindeutige Veränderungen der Kopienzahl detektiert. Dabei musste genabhängig eine bestimmte Anzahl an Exons eine veränderte Kopienzahl aufweisen, um als CNV zu gelten. Der Vergleich dieser Ergebnisse mit dem Resultat einer CNV-Bestimmung mit der qPCR-Methode (Schaeffeler et al., 2003) erbrachte eine hohe Übereinstimmung der Ergebnisse (98%). Es traten dennoch zwei Arten von Diskrepanzen auf. In machen Proben mit Duplikation war die genaue Anzahl der duplizierten Sequenzen nicht korrekt (*CYP2E1*, *SULT1A1*). Bei anderen detektierte eine der Methoden, d.h. NGS vs. qPCR eine Kopienzahl, die in der anderen nicht gefunden wurde (*CYP2D6*, *SULT1A1*). Da die TaqMan CNV-Assays nur eine ungefähr 100-200bp lange Region amplifizieren, kann mit dieser Methode nur über diesen Bereich eine Aussage zur Kopienzahl getroffen werden. Deswegen wurde zur besseren Vergleichbarkeit die Abdeckungsanalyse für jedes Exon wiederholt. Diese Resultate bestätigten alle Ergebnisse der Analyse auf Genebene, erklärten teilweise die inkonsistenten Ergebnisse und zeigten weitere, nur einzelne Exon-betreffende CNVs in *CYP2A6* und *CYP2D6,* die im folgenden Abschnitt diskutiert werden. Insgesamt konnten mit der panelbasierten Exonsequenzierung die pharmakogenetisch wichtigsten CNVs und die relative Kopienzahl der Gene *CY2A6, CYP2D6* und *SULT1A1* sowie von Genen der GST- und UGT Familie gleichzeitig und in hoher Auflösung bestimmt werden.

4.1.4 *CYP2A6* und *CYP2D6* Hybridallele

Mit der CNV-Analyse jedes Exons und den SNP-Informationen, bestehend aus Genotyp und VAF, wurden neue und bereits bekannte *CYP2D6* und *CYP2A6* Hybridallele in den IKP148-Leberproben detektiert und mit Allelen aus verschiedenen vorausgegangenen Studien verglichen (Haberl et al., 2005; Raimundo et al., 2000; Schröder et al., 2013; Toscano et al., 2006; Zanger et al., 2001). Die Ergebnisse wurden mit annotierten Allelen der humanen CYP-Allel-Nomenklatur-Datenbank (http://www.cypalleles.ki.se/; August 2015) abgeglichen.

4.1.4.1 *CYP2D6-CYP2D7* Hybridallele

Insgesamt wurde mit diesem Ansatz in 20% der Proben der *CYP2D6* Genotyp aktualisiert, wobei in der Hälfte der Fälle ein nicht-funktionelles Hybridallel (*CYP2D6*68*) im Tandem mit einem *CYP2D6*4* Allel (Gaedigk et al., 2012) ein einzelnes *CYP2D6*4* Allel ersetzte. Des Weiteren wurden nicht-funktionelle Allele wie die Tandemduplikationen eines *CYP2D6*4* Allels beobachtet bei der zusätzlich das Exon neun ausgetauscht sein kann (Gaedigk et al., 2006; Kramer et al., 2009) oder die Hybridallele *CYP2D6*66 und *CYP2D6*77+*2*, die unter *CYP2D6*13* zusammengefasst werden und sich nur in der Region, in der der Austausch zwischen den beiden Genen stattfindet, unterscheiden (Gaedigk et al., 2010). Die diskrepanten Ergebnisse zur qPCR reduzierten sich mit dieser Analyse auf drei Proben. In diesen wiesen alle Exons eine veränderte Kopienzahl auf (Abbildung 18). Da SNP-Genotypen und VAF in diesen Proben ebenfalls zum CNV-Typ passten, wird von einem falsch-negativen Ergebnis der qPCR ausgegangen. Während Probe #087 einige SNPs in Exon neun trägt, die eventuell die Bindung der Sonde und somit die Amplifikationeffizienz beeinträchtigen, tragen die anderen zwei Proben mit inkonsistenten Ergebnis (#166, #212) keine SNPs in Exon neun. Dass das für die qPCR-Methode wichtige Referenzgen in diesen zwei Proben eine von zwei abweichende Kopienzahl aufweist, kann durch die Ergebnisse weiterer CNV-Assays und den konsistenten Ergebnissen der NGS und qPCR-Analysen für andere Gene in diesen Proben ausgeschlossen werden. Als sonstige Fehlerquellen kommen Pipettierfehler, Verunreinigungen in Frage. Allerdings muss auch auf Limitationen der NGS-Methode hingewiesen werden, denn die Kopienzahlbestimmung war in nahezu der Hälfte der *CYP2D6* und *CYP2D7* Exons nicht zuverlässig oder nicht möglich. Obwohl die Sequenzierreaktion 100bp umfasste, kann diese Länge auf Grund der starken Homologie zwischen *CYP2D6* und *CYP2D7* für eine spezifische Zuordnung der Rohreads an das Referenzgenoms zu gering sein (Drögemöller et al., 2013). Die Ergebnisse sollten mit einer NGS-Methode, die längere Sequenzen lesen kann (http://www.nanoporetech.com/) oder mit weiteren Methoden, wie zum Beispiel einer

herkömmlichen Sequenzierung nach Sanger oder weiteren TaqMan CNV-Assays, überprüft werden, was in der klinischen Diagnostik für weitere homologe Gene auch empfohlen wird (Mandelker et al., 2014)

4.1.4.2 *CYP2A6-CYP2A7* Hybridallele

Wie bei der Auswertung des *CYP2D6* Locus beschrieben, war die Exonsequenzierung wie auch die Kopienzahldetektion durch homologe Sequenzen zwischen *CYP2A6* und *CYP2A7* in Exon neun und der 3'UTR gestört. Trotzdem konnte das bekannte Deletionsallel *CYP2A6*4* (Ariyoshi et al., 2002; Mwenifumbo et al., 2008), die bekannten Duplikationen *CYP2A6*1x2A* oder *B* (Fukami et al., 2007; Rao et al., 2000) und das beschriebene Hybridallel *CYP2A6*12* (Haberl et al., 2005) festgestellt werden. Die angenommene Duplikation eines *CYP2A6*12* Allels in einer Probe (#295) würde bei Bestätigung eine neue Duplikationsvariante bedeuten, was jedoch im Rahmen dieser Arbeit nicht mehr mit einer weiteren Methode bestätigt wurde.

4.1.5 Vergleich der CNV-Verteilung in ADME-Genen und den drei Kohorten

In allen drei Datensätzen trugen Proben mit afrikanischer Herkunft am meisten ADME-CNVs, gefolgt von Proben mit asiatischer und europäischer Abstammung. Allerdings waren die Unterschiede im Gegensatz zu genomweiten CNV-Detektionen nicht signifikant (Conrad et al., 2010). Durchschnittlich trug eine Person 3,8 ± 0,16 CNVs der 340 untersuchten ADME-Gene. Im Vergleich zu genomweiten Analysen, die je nach Methode zwischen 50-400 CNVs (Conrad et al., 2010; The 1000 Genomes Project Consortium, 2010) pro Person fanden, haben CNVs von ADME-Genen also ungefähr einen 1-10%igen Anteil. Allerdings waren nicht alle ADME-Gengruppen gleich betroffen. Die meisten und die häufigsten CNVs waren in der Gruppe der Phase I und II Gene zu finden. Häufige CNVs von Transportern und Modifizierern waren nur in Krebszellen zu finden. Über den Grund des Ungleichgewichts kann nur spekuliert werden. Da Transkriptionsfaktoren eine Vielzahl von Genen regulieren (trans-Effekt), haben CNVs dieser Gene eine größere Auswirkung und könnten deshalb unter negativer Selektion stehen. Der negative Selektionsfaktor für CNVs in Transportern könnte deren überwiegend ubiquitäre Expression und die wichtige endogene Rolle im Influx und Efflux verschiedenster exogener und endogener Substanzen sein. Die Schnittmenge der Ergebnisse aller Datensätze war mit neun Genen nicht allzu groß, was allerdings beim Vergleich der Resultate unterschiedlichster CNV-Bestimmungsmethoden durch die erwähnten Eigenheiten der Methoden und Algorithmen nicht anders zu erwarten war (Alkan et al., 2011). Trotzdem wurden alle pharmakogenetisch wichtigen ADME-CNVs,

wie von *CYP2A6*, *CYP2D6*, *GSTM1*, *GSTT*, *UGT2B17* und *SULT1A1* mit der NGS-Analyse in den Leberproben der Studie IKP148 und mindestens in einer der anderen Kohorte gefunden (Tabelle 9 ;He et al., 2011). Eine detailliertere Diskussion zu den einzelnen CNVs und Genen findet im Abschnitt 4.2 statt.

4.2 Funktionelle Charakterisierung der gefunden ADME-CNVs

Um den Einfluss der CNVs auf die Genexpression zu bestimmen, wurde in den drei Kohorten die Assoziation zwischen der oben beschriebenen Kopienzahl und der mRNA Expression berechnet. Die Expression lag für die meisten der von CNVs betroffenen ADME-Gene aus Mikrochipexperimenten (IKP148: Schröder et al. 2013; LCLs der HapMap-Proben: Stranger et al., 2007) oder aus RNA Sequenzierungen (TCGA: http://cancergenome.nih.gov/) vor. Für CYP2A6 und CYP2D6 wurde zusätzlich auf Aktivitätsdaten, die mit einem Substrat und per Massenspektroskopie gemessen wurden (Feidt et al., 2010; Gomes et al., 2009), zurückgegriffen. Die Assoziationsanalyse bestätigte drei unterschiedliche Verhalten der mRNA Expression auf eine veränderte Kopienzahl. Neben Genen, deren mRNA Expression mit der Kopienzahl korrelierte (Gendosis-sensitiv) kamen Kandidaten, deren mRNA scheinbar unabhängig der Kopienzahl exprimiert war (Gendosis-insensitiv) vor. Zudem wurde eine inverse Korrelation beobachtet bei welcher die Expression mit steigender Kopienzahl abnahm (Gendosis-umgekehrt). Diese verschiedenen Expressionsverhalten wurden, wie in der Einleitung bereits beschrieben, in *drosophila melanogaster* (Zhou et al., 2011), in Mais (Guo et al., 1996), für Gene auf dem Chromosom 21 in Patienten mit Trisomie 21 (Aït Yahya-Graison et al., 2007) und in gesunden humanen HapMap-Proben (Woodwark and Bateman, 2011) beschrieben. Auf das insensitive oder umgekehrte Verhalten, welches über verschiedene Dosiskompensationsmechanismen, wie gelinkte genetische Varianten, Epigenetik, monoallelische Expression oder negative Rückkopplungskreise erklärt werden kann (Veitia et al., 2008; Woodwark and Bateman, 2011), wird im Folgenden näher eingegangen. In den Tumorproben wurde in dieser Arbeit für 35% der ADME-Gene eine positive Korrelation mit CNVs bestimmt. Dieser Wert ist weit geringer als die kürzlich in Tumoren postulierten 99% (Fehrmann et al., 2015). Allerdings wurde dort der Einfluss der CNVs auf die Expression mit einer auf diese Arbeit nicht übertragbaren Assoziationanalyse bestimmt. Deren Ansatz schloss weitere Regulationsfaktoren (transkriptionelle Komponenten) in die Analyse als weitere Variable ein. Unter den Faktoren finden sich auch einige der oben genannten Faktoren, wie Hormone, Transkriptionsfaktoren, physiologische Faktoren und weitere genetische Varianten und andere Stimuli (Fehrmann et al., 2015).

4.2.1 Dosis-sensitive Gene

4.2.1.1 *GSTM1* und *GSTT1*

Die Gene *GSTM1* und *GSTT1* sind hoch polymorph und homozygote Deletionen mit einer Häufigkeit von jeweils 50% und 30% in Personen mit europäischer Abstammung bestätigten frühere Arbeiten diesbezüglich (Rose-Zerilli et al., 2009). Obwohl beide Gendeletionen durch homologe Rekombination entstehen (Sprenger et al., 2000; Xu et al., 1998), ist die Frequenz von Duplikationen viel seltener (McLellan et al., 1997). Die in DGV in den HapMap-Proben beschriebenen Duplikationen stellten sich beim Vergleich aller CNV-Daten als falsch-positive Treffer heraus. Somit wurde in keiner Probe eine Duplikation der beiden GSTs beobachtet. Die Häufigkeit der Deletionen unterschied sich weltweit. In Personen mit afrikanischer Abstammung traten sie seltener auf als in Personen mit europäischer und asiatischer Abstammung. In allen Kohorten war die mRNA Expression beider GSTs signifikant von der Kopienzahl abhängig (Abbildung 27), was in extrahepatischen Geweben wie der Lunge bereits für mRNA (Butler et al., 2011), Protein (Cantlay et al., 1994) und für Enzymaktivitätsmessungen mit spezifischen Substraten gezeigt worden war (McLellan et al., 1997; Seidegård and Ekström, 1997). In den IKP148-Leberproben erklärte die Kopienzahl 74% bzw. 38% (Pseudo-R^2 einer Medianregression) der GSTM1 oder GSTT1 mRNA Expressionsvariabilität, wobei die zur Verfügung stehenden nicht-genetischen Faktoren, wie z.B. das Geschlecht, das Alter, der Alkohol- und Nikotinkonsum, Leberfunktionsparameter (CRP, GGT) oder die Diagnose, die zur Operation führte, keinen Einfluss auf die Expression zeigten. In den HapMap-Proben mit einem Genverlust konnten zwei Expressionsmuster beobachten werden, die zu einer heterozygoten und homozygoten Deletion passen würden. Dieses Beispiel verdeutlicht, dass in DGV die Information zur Kopienzahl überwiegend hinsichtlich der Art (Verlust oder Gewinn), jedoch nicht für alle Studien die genaue oder relative Kopienzahl verfügbar ist (Abbildung 27). Im TCGA-Krebsgewebe wurde für beide GSTs Isoformen in einem kleinen Teil der Proben eine der Kopienzahl untypische Expression beobachtet. Entweder war in diesen Proben die Kopienzahlbestimmung fehlerhaft oder die Regulation der Genexpression drastisch verändert. Zusätzlich wurde in den IKP148-Leberproben bestätigt, dass die *GSTM1* CNVs keinen Einfluss auf die Expression des homologen Nachbargens *GSTM2* nehmen (Butler et al., 2011). Dies ist wichtig, da GSTM2 im Falle einer *GSTM1* Deletion scheinbar die fehlende Funktion übernehmen kann, weshalb Assoziationstudien hinsichtlich Krebs bisher keine konsistenten Daten lieferten (Bhattacharjee et al., 2013). Nach meinem Wissen gab es bisher keine Studie, die einen eventuellen Nutzen bzw.

selektiven Vorteil von GST-Deletionen untersucht hat, was wegen ihres häufigen Vorkommens nicht abwegig erscheint.

4.2.1.2 CYP2D6

Wegen der oben genannten technischen Limitation der Detektionsmethoden wurden in gesunden TCGA-Proben und DGV keine CNVs des *CYP2D6* Locus gefunden. In den TCGA-Lebertumorproben, in denen die CNV-Segmente längere Regionen umspannten und deswegen leichter detektierbar waren, konnte eine signifikante Assoziation zwischen *CYP2D6* CNV und mRNA Expression gefunden werden, was für Tumorzellen einen *CYP2D6* genotypabhängigen Arzneimittelmetabolismus nahe legt. In den Leberproben konnte mit den Resultaten der NGS-Analyse auf Genlevel die Korrelation der Kopienzahl mit der mRNA Expression bestätigt werden (Abbildung 27; Zanger et al., 2005). Mit der Feinkartierung der *CYP2D6* Allele konnten 52% (lineare Regression) der Variabilität der Enzymaktivität erklärt werden (Abbildung 29), was den enormen Einfluss der genetischen Determinante bestätigt und unterstreicht (Ingelman-Sundberg et al., 2007).

4.2.1.3 CYP2A6

In den HapMap- und gesunden TCGA-Leberproben wurden zwar *CYP2A6* CNVs gefunden, allerdings waren für alle dieser Proben mit *CYP2A6* CNV keine mRNA Expressionsdaten verfügbar. In den IKP148-Leberproben bestätigte die Assoziationsanalyse den signifikanten Einfluss ($p<0,03$) der Kopienzahl auf die mRNA Expression (Haberl et al., 2005). Die signifikante Korrelation der Genotypen, die mit der Feinkartierung bestimmt wurden, betonte den Einfluss des Genotyps auf die CYP2A6 Aktivität (Mwenifumbo and Tyndale, 2007). Zwar konnten nur 11% (lineare Regression) der Variabilität durch den Genotyp erklärt werden, doch war dieser Anteil, neben den nichtgenetischen Faktoren Geschlecht ($R^2=3\%$) und Entzündungsmarker CRP ($R^2=6\%$), die in den Leberproben einen signifikanten Einfluss auf die CYP2A6 Aktivität zeigten, am höchsten.

4.2.1.4 UGT2B17

UGT2B17 gehört wie die Gene der GST Familie zu den am häufigsten deletierten Gene im humanen Genom (McCarroll et al., 2006). Deletionen wurden auch in dieser Analyse in allen untersuchten Kohorten mit Frequenzen über 50% beobachtet, wobei sich Unterschiede zwischen den untersuchten Ethnien feststellen ließen (Abbildung 11). Es bestätigte sich, dass Personen mit afrikanischer Herkunft weniger Deletionen als Personen mit europäischer Abstammung trugen. Am häufigsten waren Deletionen

in HapMap-Proben mit asiatischer Herkunft zu beobachten. Es wird angenommen, dass diese weltweite Verteilung durch die jeweilige Lebensweise und den Metabolismus von Xenobiotika und die endogenen Rolle von UGT2B17 im Steroidhormonmetabolismus zustande kam (Xue et al., 2008). Unterstützt wird diese Hypothese durch den beobachteten signifikanten Einfluss der *UGT2B17* CNVs auf die mRNA Expression, der in den LCLs der HapMap-Proben bereits beschrieben war (McCarroll et al., 2006) und durch die Ergebnisse in humanen Leberproben (IKP148) in dieser Arbeit und weiteren Studien bestätigt werden konnte (Gallagher et al., 2010). HapMap-Proben mit einer Duplikation hatten eine vergleichbare mRNA Expression wie Proben mit zwei Kopien. Eine unabhängige Studie beschrieb zwar Duplikationen des *UGT2B17* Gens, untersuchte aber nicht den funktionellen Einfluss (Gaedigk et al., 2012). Nach den Erfahrungen im GST Locus könnte es sich bei den Duplikationen in den HapMap-Proben auch um falsch-positive Ergebnisse handeln. Das Fehlen der Expressionswerte für UGT2B17 im TCGA-Datensatz kann auf die Homologie des UGT2 Locus zurückgeführt werden, wie kürzlich mittels RNA-Seq Analyse gezeigt wurde (Tourancheau et al., 2015). Die UGT2B15 Enzymaktivität wird von der Gendosis beeinflusst (Jakobsson et al., 2006). So ist der Testosteronmetabolismus in Personen, die eine Deletion tragen, vermindert. Glucuronidiertes Testosteron (TG) wird über den Urin ausgeschieden und Personen mit homozygoter Deletion haben im Vergleich zu Personen mit zwei Genkopien einen 40% niedrigeren TG Spiegel. Deswegen wurde vorgeschlagen, die *UGT2B17* Kopienzahl bei Dopingtests zu berücksichtigen (Schulze et al., 2008).

4.2.2 Dosis-insensitive Gene

4.2.2.1 CES1

CES1 ist im Gencluster der Carboxylesterasen 1-4 das einzige Gen, das von CNVs betroffen ist (MacDonald et al., 2014; Sanghani et al., 2009). Bisher sind nur heterozygote und homozygote Duplikationen des Gens beschrieben, was durch diese Arbeit in gesundem Gewebe bestätigt wurde. Die mRNA Expression unterschied sich in Proben mit drei oder vier Kopien nicht von der Expression in Leberspendern mit zwei Genkopien (IKP148, TCGA). Dieses Verhalten wurde auch in anderen Studien festgestellt (Friedrichsen et al., 2013) und resultiert aus der Struktur der Genduplikationen. So unterscheidet sich das „Muttergen“, das als *CES1A1* bezeichnet wird, von der Tochterkopie (*CES1A2*) unter anderem durch genetische Varianten in der Promoterregion, die die transkriptionelle Aktivität der *CES1A2* Kopien um bis zu 98% reduzieren (Fukami et al., 2008; Hosokawa et al., 2008). Auch in Tumorproben zeigte die Genduplikation keinen Einfluss auf die Expression. Allerdings wurde in den malignen Proben eine Gen-

deletion beobachtet, die mit einer reduzierten mRNA Expression assoziiert war. Dies kann die Annahme unterstützen, dass *CES1A1* hauptsächlich auf transkriptioneller Ebene reguliert wird (Yang et al., 2009) und somit eine Kopie für eine normales Expressionslevel nicht ausreichend ist. Vor kurzem wurde eine Assoziation der Duplikation zu Adipositas und zu Stoffwechselparametern beschrieben (Friedrichsen et al., 2013). Die bestehende Diskrepanz zwischen der dosis-insensitiven mRNA Expression und einem *CES1* Duplikationsphänotyp könnte durch gelinkte Varianten erklärt werden, die zwar die Aktivität beeinflussen, nicht aber die Expression von CES1. Eine andere Möglichkeit wäre, dass das Pseudogen *CES1A3*, das in den meisten CES1 Haplotypen vorkommt, in der post-transkriptionellen Regulation der CES1A1/2 mRNA beteiligt ist (Piehler et al., 2008; Pink et al., 2011). Letztendlich können das dosis-insensitive Verhalten und die Dosiskompensationsmechanismen auch gewebespezifisch sein.

4.2.2.2 CYP21A2

Das adrenogenitale Syndrom ist eine autosomal-rezessiv vererbte Stoffwechselerkrankung und basiert auf einem Verlust oder einer stark reduzierten Aktivität eines der beteiligten Enzyme in der Cortisolsynthese. In 90-95% der Fälle ist das Gen *CYP21A2*, das für das Enzym Steroid 21-hydroxylase kodiert, von genetischen Varianten betroffen (White and Speiser, 2000). Neben der Cortisolproduktion kann auch die Aldosteronbiosynthese und damit der Testosteronspiegel beeinträchtigt sein. Die Auswirkungen werden in einen klassischen und einen nicht-klassischen Verlauf eingeteilt. Durch die Überproduktion von männlichen Geschlechtshormonen sind Symptome, wie eine Fehlentwicklung der primären und sekundären Geschlechtsmerkmale und eine Virilisierung zu beobachten. Im schlimmsten Fall kann ein übermäßiger Salzverlust durch zu wenig Aldosteron hinzukommen. Häufige homologe Rekombinationen und Genkonversionen zwischen *CYP21A2* und dem homologen Pseudogen *CYP21A2P1* und die Lokalisation in einem Haupthistokompatibilitätskomplex auf Chr6p21.3 (>98% identische Nukleotide) begünstigen Mutationen wie SNPs, Hybridallele und strukturelle Varianten (Krone et al., 2000). So finden sich unter den 202 beschriebenen Allelen in der Alleldatenbank (http://www.cypalleles.ki.se/cyp21.htm; August 2015) auch viele nichtfunktionelle Varianten. Die in den IKP148-Leberproben gefundene hohe Expressionsvariabilität (4,2-fach log2) könnte den genetischen Einfluss unterstreichen. Es muss allerdings angemerkt werden, dass bisher keine hohe mRNA Expression in Lebergewebe beschrieben wurde. Ob es sich tatsächlich um eine ektopische mRNA Expression oder um eine fehlerhafte Messung durch den Mikrochip handelt, wurde im Rahmen dieser Arbeit nicht geklärt. Nichtsdestotrotz war in Proben mit einer Deletionen eine tendenziell erniedrigte Expression zu beobachten. Proben mit drei Kopien zeigten kei-

ne erhöhte mRNA Expression (Abbildung 28). Bisher wurden nur duplizierte Genkopien beschrieben, die entweder zum Teil oder komplett aus dem Pseudogen bestanden oder nonsense Mutationen (z.B. rs7755898; Q318Stop) trugen (Parajes et al., 2008). Das heißt, die vorkommenden Duplikationen sind bereits mit ihrer Entstehung nichtfunktionell. Weil die Sequenzhomologie auch in dieser Genregion die Auswertung erschwerte, konnte die Vielzahl an Allelen nicht systematisch ausgewertet werden. Gerade weil der molekulare Genotyp gut mit der Ausprägung des adrenogenitalen Syndroms korreliert (Jääskeläinen et al., 1997), kann mit dem panelbasierten Vorgehen - kleine Optimierungen für homologe Bereiche vorausgesetzt- eine Vielzahl von relevanten Allelen genotypisiert und eine Vorhersage zur möglichen Ausprägung gemacht werden.

4.2.2.3 UGT2B15

Das Gen *UGT2B15* befindet sich im *UGT2B17* Locus, ist aber nicht von den häufig vorkommenden Deletionen betroffen (Ménard et al., 2009). In den gesunden Leberproben (IKP148, TCGA) wurde in jeweils einer Probe eine Deletion sowie eine Duplikation detektiert. Die geringe Probenanzahl ließ keine valide Aussage hinsichtlich des Einflusses auf die Genexpression zu. Zumindest eine weitere Studie hatte eine Deletion von *UGT2B15* beschrieben aber ebenso nicht näher funktionell untersucht (Gaedigk et al., 2012). Im Krebsgewebe wurde ein signifikanter CNV-Einfluss auf die Expression festgestellt. Weitere Studien mit größerer Probenanzahl müssen in Zukunft die Funktionalität dieser Varianten untersuchen.

4.2.2.4 UGT2B28

Deletionen des Gens *UGT2B28* sind im Vergleich zu den Deletionen des Gens *UGT2B17* mit 30% ähnlich häufig in Personen mit europäischer Abstammung. Interessanterweise war im Vergleich zu den *UGT2B17* Deletionen die weltweite Verteilung der genetischen Vielfalt in Personen mit afrikanischer Herkunft der Hypothese –Afrikaner tragen mehr Varianten als Asiaten- entsprechend. Die Häufigkeit der CNVs war in Personen mit afrikanischer Herkunft am höchsten und nahm von Personen mit europäischer Herkunft zu Personen mit asiatischer Herkunft ab. Das Vorkommen von Duplikationen in den gesunden Proben der DGV- und TCGA-Kohorte muss wie in der *UGT2B17* Genregion kritisch betrachtet werden und könnte falsch-positive Ergebnisse darstellen. In keiner Kohorte wurde eine signifikante Assoziation zwischen der Kopienzahl und mRNA Expression gefunden. Dazu muss angemerkt werden, dass obwohl in allen Kohorten mRNA Expressionswerte verfügbar waren, die Expression entweder sehr schwach (LCLs der HapMap-Proben oder gesundes und malignes TCGA-

Lebergewebe) oder die Sonde des Mikrochips für *UGT2B28* auch die mRNA Isoformen von UGT2B10 und 11 detektieren kann (IKP148). Das könnte erklären, warum in IKP148-Leberproben, trotz homozygoter Deletion, eine Expression festzustellen war (Abbildung 28). Laut Proteinatlas und weiteren Studien ist UGT2B28 nicht oder nur schwach in der humanen Leber oder Lymphoblasten exprimiert (Ohno and Nakajin, 2011; Uhlén et al., 2015). Da allerdings *UGT2B28* Deletionen zu einer Art der Nebennierenrindeninsuffizienz (Brønstad et al., 2011) und zur Tumorprogression von kolorektalem Karzinom gelinkt wurden (Angstadt et al., 2013), könnte die mRNA Expression gewebespezifisch von der Kopienzahl abhängig sein. Um den genetischen Einfluss final aufklären zu können, sollte mit einer spezifischen Methode und in einem Gewebe mit hoher Expression, wie zum Beispiel Proben aus die Gallenblase oder dem Gastrointestinaltrakt, die UGT2B28 Phänotypen gemessen und in Relation zur Kopienzahl gesetzt werden.

4.3 Pharmakogenetische Analyse von *CYP2E1* Polymorphismen

4.3.1 *CYP2E1*, ein Gendosis- insensitives Gen

Die *CYP2E1* Genkopienzahl, die mit zwei unabhängigen Methoden konsistent bestimmt worden war, korrelierte überraschenderweise in den Leberproben der IKP148-Studie weder mit mRNA Expressionsdaten, die sowohl mit einem Expressions-Mikrochip als auch per TaqMan bestimmt worden waren, noch mit der Proteinexpression oder der Enzymaktivität. Dieses Ergebnis wurde in gesundem Lebergewebe des TCGA-Projektes und in mRNA Expressionsdaten aus LCLs des HapMap-Projektes bestätigt. Während in TCGA-Tumorproben Deletionen mit leicht erniedrigter Expression einhergingen, zeigten demografische und klinische Faktoren, wie z.B. das Geschlecht, Alter oder die Lebererkrankung in der Studie IKP148 keinen Einfluss und die CYP2E1 Phänotypen. Damit konnte ausgeschlossen werden, dass diese nichtgenetischen Faktoren die Ursache der CNV-unabhängigen Expression sind. Zusammengefasst kann *CYP2E1* in gesundem Gewebe zu den sogenannten dosisinsensitiven Genen, deren Expression unabhängig von der Genkopienzahl ist, zugeordnet werden. In weiteren Analysen wurden in Proben mit europäischer Abstammung zur CYP2E1 Duplikation gekoppelte SNPs gefunden, die möglicherweise eine Erklärung für die Beobachtung bieten können.

Mit den verwendeten CNV-Detektionsmethoden konnten keine exakten Bruchpunkte der CNVs identifiziert werden, allerdings zeigten alle *CYP2E1* Exons in allen Proben

der NGS-Analyse eine veränderte Kopienzahl auf. Da dies durch beide TaqMan Assays (in der Promoter und Exon 5 Region) bestätigt wurde, konnte von Variationen ausgegangen werden, die das ganze Gen und den Promoterbereich betreffen. Diese Annahme wurde durch die in DGV beschriebenen Varianten in einer kürzlich veröffentlichten Studie unterstützt (MacDonald et al., 2014). Martis und Kollegen (2014) identifizierten eine *CYP2E1* Duplikation, deren Bruchpunkte nahe einer SD (mit einer 98%igen Sequenzgleichheit) lagen, die weit vor und hinter *CYP2E1* lokalisiert ist. Dadurch ist die komplette *CYP2E1* Genregion sowie die Gene *SCART1*, *SYCE1* und *SPRN1* in der strukturellen Variante eingeschlossen (Martis et al., 2013).

Die *CYP2E1* Duplikation trat sowohl in den IKP148-Leberproben als auch in den DGV gelisteten Studien relativ selten auf (MAF< 5%). Im Vergleich zur Duplikation war das Vorkommen der Deletion weitaus seltener, was durch einen negativen Selektionsdruck erklärt werden kann. So wurde eine Assoziation zwischen einer Mikrodeletionen der Region 10q26.3 und einer Ovarialinsuffizienz beobachtet (McGuire et al., 2011). Verantwortlich könnte das ebenfalls in der Deletion liegende Nachbargen *SYCE1* sein, welches in der Prophase der Meiose die Anordnung der homologen Chromatiden zum synaptonemalen Komplex unterstützt. So sind beide Geschlechter einer *SYCE1* Knock-out Maus infertil (Bolcun-Filas et al., 2009). Aber auch die endogene physiologische Rolle des CYP2E1 kann die Seltenheit der Deletion und die funktionelle Konservierung erklären. In Extremsituation wie längerem Nahrungsmangel kann der CYP2E1 Beitrag zur Glukoneogenese den entscheidenden evolutiven Vorteil bedeuten. In Tumorproben, in denen die Deletion häufiger auftrat, könnten dagegen die Vorteile einer Gendeletion hinsichtlich der CYP2E1 Rolle in der ROS Produktion und Toxizität überwiegen.

Die Existenz von SNPs, die zur *CYP2E1* Duplikation gekoppelt sind, könnte eine Erklärung für das dosis-insensitive Verhalten in Personen europäischer Abstammung liefern. Dazu gehören SNPs, die Bindestellen von Transkriptionsfaktoren in der Promoterregion und die Regulation des Chromatinzustands beeinflussen können. Vielversprechend war die Vorhersage für den SNP in der 3'UTR (rs7081484, C>T), der miRNA Bindestellen beeinflusst. Allerdings fehlen bisher detaillierte Informationen zu den vorhergesagten miRNAs miR-4450, miR-4525, miR-5010-5p und miR-506-5p. Allerdings können andere Mechanismen der Dosis-Kompensation, wie z.B. epigenetische Faktoren, eine monoallelische Expression oder ein stochastisches Regulationsmodell, nicht ausgeschlossen werden. Es ist außerdem möglich, dass die posttranslationale und endokrine Regulation der CYP2E1 Expression die genetische Variabilität übersteigt und somit der CNV-Effekt überlagert wird. All diese Hypothesen können besonders in anderen Ethni-

zitäten wie z.B. in Proben mit afrikanischer Abstammung, in denen keine gelinkten SNPs gefunden wurden, relefant sein.

Die Dosis-Insensivität von CYP2E1 hat Implikationen für genetische Assoziationsanalysen. Vor kurzem wurde eine Assoziation zwischen 10q26.3 Duplikationen und Adipositas in US-Amerikanern europäischer und afrikanischer Abstammung beschrieben (Yang et al., 2013). Die Daten in dieser Arbeit lassen vermuten, dass diese Assoziation und mögliche weitere klinische Korrelationen mit *CYP2E1* CNVs nicht auf eine CYP2E1 Überexpression rückzuführen sind, sondern durch andere Gene innerhalb der Duplikation oder gelinkte genetische Varianten zustande kommt. Eine ähnliche Beobachtung wurde wie oben in 4.2.2.1 beschrieben für *CES1* gemacht.

Die Daten aus der IKP148-Leberkohorte wurden auch im Hinblick auf die CYP2E1 Populationsvariabilität ausgewertet. In den Leberproben wurde im Vergleich zu einer kürzlich veröffentlichten Metaanalyse zum CYP2E1 Enzymvorkommen eine größere interindividuelle Variabilität (168-fach) der Proteinexpression beobachtet (Achour et al., 2014). Ferner konnte eine Korrelation zwischen der Protein und mRNA Levels festgestellt werden (r_S=0.48, p<10-9), die in anderen Studien nicht beobachtet wurde (Ohtsuki et al., 2012). Diese Abweichung kann durch die Art und Anzahl der verwendeten Leberproben oder durch die Quantifizierungsmethode entstehen. Im Gegensatz zu den anderen Studien wurde in dieser Arbeit ein hochspezifischer monoklonaler Antikörper verwendet (Gelboin et al., 1996).

CYP2E1 ist ein durch Alkohol induzierbares Enzym (Ingelman-Sundberg et al., 1993). Das unerwartete Fehlen einer Korrelation zwischen CYP2E1 Phänotypen und dem Alkoholkonsum kann durch die Tatsache erklärt werden, dass die verwendeten Leberproben durch einen operativen Eingriff gewonnen wurden. Vor solch einem Eingriff müssen Patienten mindestens ein bis zwei Tage auf Alkohol verzichten und nüchtern erscheinen. Durch die relativ kurze Halbwertszeit des CYP2E1 Proteins kehren durch Alkohol induzierte CYP2E1 Proteinlevel daher schnell auf Normalniveau zurück (Gonzalez, 2007; Oneta et al., 2002). Andere nicht-genetische Faktoren zeigten wenig Einfluss auf die CYP2E1 Phänotypen. Nur die CYP2E1 Proteinexpression war in der multivariaten Analyse geschlechtsabhängig (in Spenderinnen) und bei Entzündungsanzeichen (hoher CRP Spiegel) signifikant erhöht. Diese Beobachtung und die hohe Variabilität der Proteinlevel unterstützen die generelle Annahme, dass posttranskriptionelle und posttranslationale Mechanismen wichtiger als die transkriptionale Aktivität für die CYP2E1 Regulation sind (Bolt et al., 2003; Novak and Woodcroft, 2000).

Die hier gemachten Untersuchungen zeigten, dass die Expression von CYP2E1 sowohl in gesundem Lebergewebe als auch in LCLs unabhängig von der Gendosis ist.

Obwohl der finale Mechanismus ungeklärt bleibt, liefern diese Daten die Grundlage für weitere Experimente. Zusammengefasst könnte die wichtige physiologische Rolle des CYP2E1 die Seltenheit der Deletion erklären. Im Gegensatz dazu würde die gefährliche Aktivität des CYP2E1 in der ROS Entstehung und der einhergehenden Toxizität ein entscheidender evolutiver Nachteil der zusätzlichen Genkopien bedeuten.

4.3.2 Assoziationsanalyse zwischen *CYP2E1* Polymorphismen und dem Schilddrüsenkrebsrisiko

Viele Fall-Kontroll-Studien untersuchten bisher die Assoziation von SNPs in *CYP2E1* mit einer Vielzahl von Krankheiten und Krebsentitäten. Allerdings beschränkten sich die meisten Studien auf die Allele *5A/B (rs2031920, rs3813867) und *6 (rs6413432) (Dey, 2013; Neafsey et al., 2009). Im Rahmen dieser Arbeit ergab sich eine interessante Kooperation mit der Gruppe von Prof. Stefano Landi von der Universität Pisa, Italien. Diese Gruppe hatte in einer DTC Fall-Kontroll-Kohorte systematisch mit einem tag-SNP Ansatz vier SNPs ausgewählt, die zu SNPs über die ganze *CYP2E1* Genregion gekoppelt waren und deren Beziehung zum Schilddrüsenkrebs untersucht. Der tag-SNP rs2480258 (G>A), der SNPs in Intron acht bis in der 3'UTR abdeckt, wies eine signifikante Assoziation zum Krebsrisiko auf. Eine funktionelle Analyse in den 150 humanen Leberproben (IKP148) zeigte, dass das zum Risiko assoziierte A Allel signifikant die mRNA und Proteinexpression wie auch die Enzymaktivität reduzierte. Die funktionelle *in silico* Analyse aller SNPs im tag-SNP Haplotyp sagte für die SNPs in der 3'UTR (rs24080257, rs2480256) veränderte Bindestellen für miRNAs voraus (miR-5582, miR-570, miR-203, miR-510). Im Gegensatz zu den oben in 4.3.1 genannten miRNAs wurde kürzlich ein negativer Einfluss der miR-570 auf die Expression von CYP2E1 in gesunden humanen Lebern beobachtet (Nakano et al., 2015). Zusammengefasst lässt sich vermuten, dass der CYP2E1 Phänotyp in Trägern des rs2480258 getaggten Haplotyps (A Allel) per miR570 herunterreguliert wird, wobei im Gegensatz dazu der häufigere Haplotyp (G Allel) nicht von der miRNA erkannt und daher nicht reguliert wird. Es muss jedoch auf Unstimmigkeiten des tag-SNP Einflusses auf die mRNA Expression hingewiesen werden. Der in dieser Arbeit gezeigte negative Einfluss des A Allels auf die mRNA Expression konnte in der Arbeit von Nakano et al. (2015) und in eQTL Daten aus Lebergewebe nicht reproduziert werden (Innocenti et al., 2011; Schadt persönliche Mitteilung). Solche Diskrepanzen können durch die unterschiedliche Herkunft der Leberproben und der unterschiedlichen mRNA Messmethode zustande kommen und müssen durch weitere Analysen geklärt werden.

Insgesamt konnte also ein Einfluss von einer *CYP2E1* Variante auf die Karzinogenese der Schilddrüse beobachtet werden. Eine ausführliche Diskussion des Zusammen-

hangs mit der Acrylamid Toxizität findet sich in der Publikation (Pellé, Cipollini, Tremmel et al., 2015). Für diesen bisher unbekannten Polymorphismus in der 3'UTR wurde die Beziehung zur Enzymaktivität gezeigt und die miR570 als molekularer Mechanismus vorgeschlagen. In der Zukunft sollte dieser Haplotyp in Arbeiten zum Acrylamid Metabolismus aber auch in anderen Fall-Kontroll-Studien mit berücksichtigt werden. Die 3'UTR-SNPs im beschriebenen Haplotyp sind nicht zu dem in 4.3.1 erwähnten Duplikation-gelinkten 3'UTR-SNP gelinkt. Dieser Fakt und die Ergebnisse dieser Arbeit legen die Basis für eine dringende systematische Analyse der *CYP2E1* 3'UTR und putativen genotypabhängigen regulierenden miRNAs.

4.4 *SULT1A1* Polymorphismen beeinflussen das Methyleugenol DNA-Adduktlevel. Gefahr für Leib und Leber?

Durch diese Analyse wurde bestätigt, dass in humaner Leber die SULT1A1 mRNA und Protein Expression durch die *SULT1A1* Kopienzahl signifikant beeinflusst wird. Eine Kooperation mit der Gruppe von Prof. Dr. Hans-Rudolf Glatt vom Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke, Abteilung Ernährungstoxikologie erlaubte eine weitere Assoziationanalyse für die *SULT1A1* CNVs. Diese Gruppe zeigte, dass die Natursubstanz Methyleugenol von SULT1A1 metabolisiert wird und das Produkt karzinogen wirken kann (Herrmann et al., 2012). Außerdem konnte die Arbeitsgruppe in 121 Leberproben (IKP148) Methyleugenol DNA-Addukte mittels Massenspektroskopie messen (Herrmann et al., 2013). Interessanterweise korrelierten die DNA-Adduktmessungen mit der *SULT1A1* Kopienzahl.

Die *SULT1A1* CNVs wurden mit zwei unabhängigen Methoden (NGS und qPCR) in den Leberproben bestimmt. Die Ergebnisse waren in drei von 121 Proben inkonsistent (Abbildung 37). Die NGS-Methode zeigte in der Analyse auf Exonebene, dass in jeder Probe mit CNV nahezu alle Exons eine veränderte Kopienzahl aufwiesen (Abbildung A2). Da die Sequenzen der Gene *SULT1A1*, *SULT1A2* und *SULT1A3* nahezu identisch sind, könnte es auch sein, dass Varianten die Homologie der Sequenzen ändern und deshalb die TaqMan Sonde im positiven Fall (SNP-bedingte höhere Sequenzgleichheit) zusätzlich eine falsche Region amplifiziert. Die CNV-Daten der NGS-Analyse zeigten, dass 36% der Proben mehr als zwei *SULT1A1* Genkopien tragen. Deletionen wurden nur in 4% der Spender gefunden. Die Frequenzen waren vergleichbar zu Ergebnissen einer früheren Studie (Gaedigk et al., 2012).

Mit einer WesternBlot Analyse wurde bestätigt, dass SULT1A1 stärker als SULT1A2 und SULT1A3 nicht in adulter humaner Leber exprimiert ist (Jancova et al., 2010). Die relative SULT1A1 Expression zeigte eine schwache inter-individuelle Variabilität und korrelierte signifikant mit der mRNA Expression. Nicht-genetische Faktoren beeinflussten die SULT1A1 Expressionsphänotypen nicht. Die Korrelation der mRNA und Proteinexpression mit den Adduktmessungen (r_s=0,43; r_s=0,44) bestätigte die Bedeutung von SULT1A1 im Methyleugenolstoffwechsel und zeigte diese erstmals in humanem Lebergewebe (Herrmann et al., 2012).

Die SULT1A1 mRNA und Proteinexpression war signifikant von der *SULT1A1* Kopienzahl abhängig. Außerdem wurde auch in TCGA-Lebertumorproben eine signifikante Assoziation festgestellt. Obwohl der Unterschied der Proteinexpression zwischen Trägern von drei und vier Kopien nicht mehr so ausgeprägt war, erklärte die strukturelle Variante fast 40% der Proteinvariabilität. Wie durch Ergebnisse in einer Sult1a1-defizienten Maus bereits spekuliert worden war (Herrmann et al., 2013), konnte in dieser Arbeit in humanem Lebergewebe gezeigt werden, dass das Methyleugenol DNA-Adduktlevel signifikant von der *SULT1A1* Genkopienzahl abhängig ist. Die CNVs erklärten ungefähr 13% der hohen Adduktvariabilität (122-fach). Ein weiterer Faktor war die in den Leberspendern beobachtete Assoziation zwischen einem hohen Spiegel des Entzündungsmarker CRP und leicht reduzierten DNA-Addukten. Da in Proben mit erhöhtem CRP das ADME System herunterreguliert ist (Klein et al., 2014), kann spekuliert werden, dass der Methyleugenolmetabolismus in Personen mit einer Entzündungsreaktion langsamer und geringer ausfällt. Überraschenderweise war in älteren Spendern ein geringerer DNA-Adduktwert zu beobachten. Allerdings war diese Korrelation nicht stark ausgeprägt (r_S= -0,26). Faktoren, die eine Erklärung liefern können, sind die Einnahme von mehr Medikamenten, die den Methyleugenolmetabolismus stören, ein geringerer hepatischer Blutdurchfluss und eine Nahrung, die weniger Methyleugenol enthält. Zudem sollte angemerkt werden, dass die meisten der Spender über 60 Jahre alt waren.

Nur die SULT1A1 mRNA Expression war vom nicht-synonymen SNP (SULT1A1*2; rs9282861; G638A; Arg213His) beeinflusst. Die Assoziation war jedoch gegenüber dem Einfluss der Kopienzahl gering und auf Proteinlevel unbedeutend, was durch die GWAS und auch von anderen Studien beobachtet wurde (Hebbring et al., 2007; Yu et al., 2010).

Die Expression und Polymorphismen weiterer ADME-Gene (CYP2A6, UGT1A1, UGT1A3, GSTM1 und GSTT1), deren Rolle im Methyleugenolmetabolismus bekannt oder spekuliert wurde, korrelierten, wie auch Komponenten des DNA-Reparatursystems, nicht mit den DNA-Adduktmessungen. Obwohl in den Leberproben

keine Beziehung zwischen den DNA-Adduktlevels oder der SULT1A1 Proteinexpression mit dem Erkrankungsrisiko an einem Leberkarzinom gefunden wurde, kann die toxikologische Rolle und die Karzinogenität von Methyleugenol gerade hinsichtlich der *SULT1A1* Kopienzahlabhängigkeit nicht ausgeschlossen werden. Träger von multiplen *SULT1A1* Genkopien (>3 Kopien, 6% in Personen europäischer Abstammung) hatten ein 2,8-fach höheres Level an DNA-Addukten im Vergleich zu Trägern einer Genkopie. Hypothetisch können deshalb Träger mehrerer *SULT1A1* Kopien viel leichter, schneller und öfter ein kritisches DNA Adduktlevel erreichen, das zu einem höheren Krebsrisiko führt. Weiterführende Studien sollten unbedingt neben der Menge des aufgenommenen Methyleugenols die SULT1A1 Kopienzahl berücksichtigen. Sehr vereinfacht und übertrieben dargestellt, sollten Personen mit mehreren *SULT1A1* Kopien vielleicht ihren Tomate-Mozzarella-Basilikum Konsum überdenken oder wenigstens Fencheltee beiseitelassen um ihr Krebsrisiko nicht unnötig zu steigern.

5 Fazit

Die kombinierte Auswertung von Daten zu CNVs aus frei zugänglichen Ressourcen (DGV und TCGA) und 150 Leberproben (IKP148), deren Genotyp und Phänotyp mit aktuellsten Methoden bestimmt wurde, hat sich als erfolgreicher Ansatz erwiesen, systematisch das Vorkommen von CNVs in ADME-Genen und deren funktionellen Einfluss auf die Expression und Aktivität der wichtigsten ADME-Gene zu bestimmen. Mit den panelbasierten Exon-NGS-Daten und dem entwickelten Algorithmus für die CNV-Detektion konnten trotz der Probleme bei hohen Sequenzhomologien CNVs der ADME-Gene in den IKP148-Leberproben zuverlässig bestimmt werden. Es zeigte sich, dass Phase I und II Gene im Gegensatz zu Transportern und Regulatoren häufiger von CNVs beeinflusst waren. Innerhalb der Phase I war die *CYP450* Genfamilie die am stärksten betroffene Gruppe. In Phase II waren sehr häufige CNVs in den Genfamilien *UGT2*, *GST*, sowie dem Gen *SULT1A1* gefunden worden. Ein möglicher Grunde für das Auftreten von CNVs speziell in diesen Gengruppen ist die gemeine Lage der Gene in DNA Regionen mit hohen Homologien und SDs, die für Rekombinationsereignisse prädisponiert sind.

Die Auswirkung der CNVs auf den Phänotyp war unterschiedlich. Einerseits war eine signifikante Assoziation zur mRNA, Proteinexpression oder Enzymaktivität für CNVs der Gene *CYP2D6, SULT1A1, GSTM1, GSTT1* und *UGT2B17* zu beobachten. Andererseits hatte die Kopienzahl keinen (Gendosis-insensitiv) oder nur teilweise einen Einfluss auf den Phänotyp. Dies wurde für die Gene *CYP2E1*, *CES1* (signifikanter Einfluss nur von den Deletionen oder nur von den Duplikationen) und *CYP21A2*, *UGT2B15* und *UGT2B28* (gewebespezifisch) beobachtet. Evolutiv betrachtet, kann die endogene Genfunktion - also die Beteiligung der ADME-Gene an lebenswichtigen Prozessen - die Etablierung von CNVs in einer Population selektieren. Das Vorkommen von häufigen CNVs (*GSTM1*, *CYP2D6*, *SULT1A1*) ist für das Überleben und die Fortpflanzung vorteilhaft bzw. unbedeutend. Im Gegensatz dazu, können sich CNVs von Genen mit relevanter endogener Funktion nur in einer Population etablieren, wenn der CNV-Effekt auf den Phänotyp sehr gering oder abwesend ist (Duplikationen der Gene *CYP2E1*, *CES1* oder *CYP21A2*).

Mechanismen, die eine veränderte Gendosis und damit eine höhere Expression kompensieren können, sind vielfältig, teilweise noch nicht komplett verstanden und gewebe- und tumorspezifisch. Unter anderem wurden gelinkte genetische Varianten, miRNAs, eine monoallelische Expression, epigenetische Faktoren (Imprinting), ein unvoll-

ständiger Einschluss der regulatorischen oder kodierenden Sequenzen, negative Feedbackschleifen, Duplikationen im Tandem inaktivieren sich gegenseitig oder stöchiometrische Genregulationsmodelle als mögliche Mechanismen diskutiert (Henrichsen et al., 2009b; Veitia et al., 2008; Woodwark and Bateman, 2011).

In dieser Arbeit wurden in Proben mit europäischer Abstammung zur Duplikation gelinkte SNPs (Promoter und 3‘UTR) als Faktoren für die Stilllegung einer *CYP2E1* Kopie postuliert. Diese Art der Dosiskompensation ist bereits für *CES1* und *CYP21A2* beschrieben. Allerdings könnte auch die sehr interindividuelle CYP2E1 mRNA und Proteinexpression, die durch eine starke post-transkriptionale und post-translationale Regulation entsteht, den CNV-Effekt verdecken bzw. kompensieren.

In Kooperation mit der Gruppe von Prof. Stefano Landi von der Universität Pisa, Italien wurde in weiteren Untersuchungen zu *CYP2E1* ein SNP in der 3’UTR von *CYP2E1* identifiziert, welcher zu einem erhöhten Risiko für Schilddrüsenkrebs und zu einer verminderten CYP2E1 Expression und Enzymaktivität assoziiert war. Die Variante liegt innerhalb einer miRNA Bindestelle und führt zu einer Änderung der miRNA Bindungsaffinität, was zu einer genotypabhängigen unterschiedlichen Regulation der CYP2E1 Expression führen kann. Dieses Ergebnis einer vom Genotyp abhängigen post-transkriptionalen Genregulation legt die Grundlage für weitere Studien und Arbeiten zum direkten molekularen Mechanismus und zu einer systematischen Untersuchungen der wichtigen 3’UTR des *CYP2E1* Gens.

Der Einfluss der *SULT1A1* CNVs auf die Genexpression war bereits bekannt und wurde durch die Ergebnisse dieser Arbeit bestätigt und durch die Proteindaten verstärkt. Darüber hinaus wurde in Kooperation mit der Gruppe von Prof. Dr. Hans-Rudolf Glatt vom Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke, Abteilung Ernährungstoxikologie gezeigt, dass SULT1A1 im humanen Methyleugenolmetabolismus eine wichtige Rolle spielt und genetische strukturelle Varianten des Gens die DNA-Adduktvariabilität signifikant beeinflussen. Träger von mehreren *SULT1A1* Genkopien hatten im Vergleich zu Trägern mit einer oder zwei Kopien hohe Adduktwerte. Weitere epidemiologische Studien müssen klären, inwieweit das Krebsrisiko von der *SULT1A1* Kopienzahl beeinflusst wird.

Literaturverzeichnis

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Achour, B., Barber, J., and Rostami-Hodjegan, A. (2014). Expression of hepatic drug-metabolizing cytochrome p450 enzymes and their intercorrelations: a meta-analysis. Drug Metab. Dispos. Biol. Fate Chem. *42*, 1349–1356.

Aitken, A.E., Richardson, T.A., and Morgan, E.T. (2006). Regulation of drug-metabolizing enzymes and transporters in inflammation. Annu. Rev. Pharmacol. Toxicol. *46*, 123–149.

Aït Yahya-Graison, E., Aubert, J., Dauphinot, L., Rivals, I., Prieur, M., Golfier, G., Rossier, J., Personnaz, L., Creau, N., Bléhaut, H., et al. (2007). Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am. J. Hum. Genet. *81*, 475–491.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. *41*, 1061–1067.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Angstadt, A.Y., Berg, A., Zhu, J., Miller, P., Hartman, T.J., Lesko, S.M., Muscat, J.E., Lazarus, P., and Gallagher, C.J. (2013). The effect of copy number variation (CNV) in the phase II detoxification genes, UGT2B17 and UGT2B28, on colorectal cancer risk. Cancer *119*, 2477–2485.

Anzenbacher, P., and Anzenbacherová, E. (2012). Drug-Metabolzining Enzymes - An Overview. In Metabolism of Drugs and Other Xenobitoics, (Weinheim, Germany: Wiley-VCH), pp. 3–25.

Ariyoshi, N., Sekine, H., Saito, K., and Kamataki, T. (2002). Characterization of a genotype previously designated as CYP2A6 D-type: CYP2A6*4B, another entire gene deletion allele of the CYP2A6 gene in Japanese. Pharmacogenetics *12*, 501–504.

Arslan, S., Silig, Y., and Pinarbasi, H. (2011). Sulfotransferase 1A1 Arg(213)His polymorphism and prostate cancer risk. Exp. Ther. Med. *2*, 1159–1162.

Beckmann, J.S., Sharp, A.J., and Antonarakis, S.E. (2008). CNVs and genetic medicine (excitement and consequences of a rediscovery). Cytogenet. Genome Res. *123*, 7–16.

Beierle, I., Meibohm, B., and Derendorf, H. (1999). Gender differences in pharmacokinetics and pharmacodynamics. Int. J. Clin. Pharmacol. Ther. *37*, 529–547.

Beland, F.A., Mellick, P.W., Olson, G.R., Mendoza, M.C.B., Marques, M.M., and Doerge, D.R. (2013). Carcinogenicity of acrylamide in B6C3F(1) mice and F344/N rats from a 2-year drinking water exposure. Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc. *51*, 149–159.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Bhattacharjee, P., Paul, S., Banerjee, M., Patra, D., Banerjee, P., Ghoshal, N., Bandyopadhyay, A., and Giri, A.K. (2013). Functional compensation of glutathione S-transferase M1 (GSTM1) null by another GST superfamily member, GSTM2. Sci. Rep. *3*, 2704.

Bolcun-Filas, E., Hall, E., Speed, R., Taggart, M., Grey, C., de Massy, B., Benavente, R., and Cooke, H.J. (2009). Mutation of the mouse Syce1 gene disrupts synapsis and suggests a link between synaptonemal complex structural components and DNA repair. PLoS Genet. *5*, e1000393.

Bolt, H.M., Roos, P.H., and Thier, R. (2003). The cytochrome P-450 isoenzyme CYP2E1 in the biological processing of industrial chemicals: consequences for occupational and environmental medicine. Int. Arch. Occup. Environ. Health *76*, 174–185.

Brønstad, I., Wolff, A.S.B., Løvås, K., Knappskog, P.M., and Husebye, E.S. (2011). Genome-wide copy number variation (CNV) in patients with autoimmune Addison's disease. BMC Med. Genet. *12*, 111.

Butler, M.W., Hackett, N.R., Salit, J., Strulovici-Barel, Y., Omberg, L., Mezey, J., and Crystal, R.G. (2011). Glutathione S-transferase copy number variation alters lung gene expression. Eur. Respir. J. Off. J. Eur. Soc. Clin. Respir. Physiol. *38*, 15–28.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*, 722–729.

Cantlay, A.M., Smith, C.A., Wallace, W.A., Yap, P.L., Lamb, D., and Harrison, D.J. (1994). Heterogeneous expression and polymorphic genotype of glutathione S-transferases in human lung. Thorax *49*, 1010–1014.

Caro, A.A., and Cederbaum, A.I. (2004). Oxidative stress, toxicology, and pharmacology of CYP2E1. Annu. Rev. Pharmacol. Toxicol. *44*, 27–42.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Cotreau, M.M., von Moltke, L.L., and Greenblatt, D.J. (2005). The influence of age and sex on the clearance of cytochrome P450 3A substrates. Clin. Pharmacokinet. *44*, 33–60.

Das, P., Shaik, A.P., and Bammidi, V.K. (2009). Meta-analysis study of glutathione-S-transferases (GSTM1, GSTP1, and GSTT1) gene polymorphisms and risk of acute myeloid leukemia. Leuk. Lymphoma *50*, 1345–1351.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Dey, A. (2013). Cytochrome P450 2E1: its clinical aspects and a brief perspective on the current research scenario. Subcell. Biochem. *67*, 1–104.

Doligalski, C.T., Tong Logan, A., and Silverman, A. (2012). Drug Interactions. Gastroenterol. Hepatol. *8*, 376–383.

Drögemöller, B.I., Wright, G.E.B., Niehaus, D.J.H., Emsley, R., and Warnich, L. (2013). Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. Pharmacogenet. Genomics *23*, 666–674.

Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R., and Sikela, J.M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res. *17*, 1266–1277.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Eichelbaum, M., Ingelman-Sundberg, M., and Evans, W.E. (2006). Pharmacogenomics and individualized drug therapy. Annu. Rev. Med. *57*, 119–137.

Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Westra, H.-J., Maloney, D., Simeonov, A., Pers, T.H., Hirschhorn, J.N., Jansen, R.C., Schultes, E.A., et al. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. *47*, 115–125.

Feidt, D.M., Klein, K., Hofmann, U., Riedmaier, S., Knobeloch, D., Thasler, W.E., Weiss, T.S., Schwab, M., and Zanger, U.M. (2010). Profiling induction of cytochrome p450 enzyme activity by statins using a new liquid chromatography-tandem mass spectrometry cocktail assay in human hepatocytes. Drug Metab. Dispos. Biol. Fate Chem. *38*, 1589–1597.

Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.

Friedrichsen, M., Poulsen, P., Wojtaszewski, J., Hansen, P.R., Vaag, A., and Rasmussen, H.B. (2013). Carboxylesterase 1 Gene Duplication and mRNA Expression in Adipose Tissue Are Linked to Obesity and Metabolic Function. PLoS ONE *8*, e56861.

Fukami, T., Nakajima, M., Yamanaka, H., Fukushima, Y., McLeod, H.L., and Yokoi, T. (2007). A novel duplication type of CYP2A6 gene in African-American population. Drug Metab. Dispos. Biol. Fate Chem. *35*, 515–520.

Fukami, T., Nakajima, M., Maruichi, T., Takahashi, S., Takamiya, M., Aoki, Y., McLeod, H.L., and Yokoi, T. (2008). Structure and characterization of human carboxylesterase 1A1, 1A2, and 1A3 genes. Pharmacogenet. Genomics *18*, 911–920.

Gaedigk, A., Blum, M., Gaedigk, R., Eichelbaum, M., and Meyer, U.A. (1991). Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. Am. J. Hum. Genet. *48*, 943–950.

Gaedigk, A., Bradford, L.D., Alander, S.W., and Leeder, J.S. (2006). CYP2D6*36 gene arrangements within the cyp2d6 locus: association of CYP2D6*36 with poor metabolizer status. Drug Metab. Dispos. Biol. Fate Chem. *34*, 563–569.

Gaedigk, A., Jaime, L.K.M., Bertino, J., Bérard, A., Pratt, V., Bradford, L.D., and Leeder, J.S. (2010). Identification of novel CYP2D7-2D6 hybrids: non-functional and functional variants. Pharmacogenetics Pharmacogenomics *1*, 121.

Gaedigk, A., Twist, G.P., and Leeder, J.S. (2012). CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. Pharmacogenomics *13*, 91–111.

Gallagher, C.J., Balliet, R.M., Sun, D., Chen, G., and Lazarus, P. (2010). Sex Differences in UDP-Glucuronosyltransferase 2B17 Expression and Activity. Drug Metab. Dispos. *38*, 2204–2209.

Gandhi, M., Aweeka, F., Greenblatt, R.M., and Blaschke, T.F. (2004). Sex differences in pharmacokinetics and pharmacodynamics. Annu. Rev. Pharmacol. Toxicol. *44*, 499–523.

Gelboin, H.V., Goldfarb, I., Krausz, K.W., Grogan, J., Korzekwa, K.R., Gonzalez, F.J., and Shou, M. (1996). Inhibitory and noninhibitory monoclonal antibodies to human cytochrome P450 2E1. Chem. Res. Toxicol. *9*, 1023–1030.

Gomes, A.M., Winter, S., Klein, K., Turpeinen, M., Schaeffeler, E., Schwab, M., and Zanger, U.M. (2009). Pharmacogenomics of human liver cytochrome P450 oxidoreductase: multifactorial analysis and impact on microsomal drug oxidation. Pharmacogenomics *10*, 579–599.

Gonzalez, F.J. (2007). The 2006 Bernard B. Brodie Award Lecture. Cyp2e1. Drug Metab. Dispos. Biol. Fate Chem. *35*, 1–8.

Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. PathoGenetics *1*, 4.

Guo, M., Davis, D., and Birchler, J.A. (1996). Dosage Effects on Gene Expression in a Maize Ploidy Series. Genetics *142*, 1349–1355.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S.A.A.C., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J.D., et al. (2008). Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. *40*, 538–545.

Haberl, M., Anwald, B., Klein, K., Weil, R., Fuss, C., Gepdiremen, A., Zanger, U.M., Meyer, U.A., and Wojnowski, L. (2005). Three haplotypes associated with CYP2A6 phenotypes in Caucasians. Pharmacogenet. Genomics *15*, 609–624.

Harewood, L., Chaignat, E., and Reymond, A. (2012). Structural variation and its effect on expression. Methods Mol. Biol. Clifton NJ *838*, 173–186.

Hayes, J.D., Flanagan, J.U., and Jowsey, I.R. (2005). Glutathione transferases. Annu. Rev. Pharmacol. Toxicol. *45*, 51–88.

He, Y., Hoskins, J.M., and McLeod, H.L. (2011). Copy number variants in pharmacogenetic genes. Trends Mol. Med.

Hebbring, S.J., Adjei, A.A., Baer, J.L., Jenkins, G.D., Zhang, J., Cunningham, J.M., Schaid, D.J., Weinshilboum, R.M., and Thibodeau, S.N. (2007). Human SULT1A1

gene: copy number differences and functional implications. Hum. Mol. Genet. *16*, 463–470.

Hebbring, S.J., Moyer, A.M., and Weinshilboum, R.M. (2009). Sulfotransferase gene copy number variation: pharmacogenetics and function. Cytogenet. Genome Res. *123*, 205–210.

Henrichsen, C.N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009a). Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. *41*, 424–429.

Henrichsen, C.N., Chaignat, E., and Reymond, A. (2009b). Copy number variants, diseases and gene expression. Hum. Mol. Genet. *18*, R1–R8.

Herrmann, K., Engst, W., Appel, K.E., Monien, B.H., and Glatt, H. (2012). Identification of human and murine sulfotransferases able to activate hydroxylated metabolites of methyleugenol to mutagens in Salmonella typhimurium and detection of associated DNA adducts using UPLC-MS/MS methods. Mutagenesis *27*, 453–462.

Herrmann, K., Schumacher, F., Engst, W., Appel, K.E., Klein, K., Zanger, U.M., and Glatt, H. (2013). Abundance of DNA adducts of methyleugenol, a rodent hepatocarcinogen, in human liver samples. Carcinogenesis.

Hoensch, H., Peters, W.H.M., Roelofs, H.M.J., and Kirch, W. (2006). Expression of the glutathione enzyme system of human colon mucosa by localisation, gender and age. Curr. Med. Res. Opin. *22*, 1075–1083.

Homer, N., Merriman, B., and Nelson, S.F. (2009). Local alignment of two-base encoded DNA sequence. BMC Bioinformatics *10*, 175.

Hosokawa, M., Furihata, T., Yaginuma, Y., Yamamoto, N., Watanabe, N., Tsukada, E., Ohhata, Y., Kobayashi, K., Satoh, T., and Chiba, K. (2008). Structural organization and characterization of the regulatory element of the human carboxylesterase (CES1A1 and CES1A2) genes. Drug Metab. Pharmacokinet. *23*, 73–84.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Huang, S.-M., and Temple, R. (2008). Is this the drug or dose for you? Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. Clin. Pharmacol. Ther. *84*, 287–294.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.

Ingelman-Sundberg, M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J. *5*, 6–13.

Ingelman-Sundberg, M., Johansson, I., Yin, H., Terelius, Y., Eliasson, E., Clot, P., and Albano, E. (1993). Ethanol-inducible cytochrome P4502E1: genetic polymorphism, regulation, and possible role in the etiology of alcohol-induced liver disease. Alcohol Fayettev. N *10*, 447–452.

Ingelman-Sundberg, M., Sim, S.C., Gomez, A., and Rodriguez-Antona, C. (2007). Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol. Ther. *116*, 496–526.

Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. *7*, e1002078.

International Transporter Consortium, Giacomini, K.M., Huang, S.-M., Tweedie, D.J., Benet, L.Z., Brouwer, K.L.R., Chu, X., Dahlin, A., Evers, R., Fischer, V., et al. (2010). Membrane transporters in drug development. Nat. Rev. Drug Discov. *9*, 215–236.

Jääskeläinen, J., Levo, A., Voutilainen, R., and Partanen, J. (1997). Population-wide evaluation of disease manifestation in relation to molecular genotype in steroid 21-hydroxylase (CYP21) deficiency: good correlation in a well defined population. J. Clin. Endocrinol. Metab. *82*, 3293–3297.

Jakobsson, J., Ekström, L., Inotsume, N., Garle, M., Lorentzon, M., Ohlsson, C., Roh, H.-K., Carlström, K., and Rane, A. (2006). Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. J. Clin. Endocrinol. Metab. *91*, 687–693.

Jancova, P., Anzenbacher, P., and Anzenbacherova, E. (2010). Phase II drug metabolizing enzymes. Biomed. Pap. Med. Fac. Univ. Palacký Olomouc Czechoslov. *154*, 103–116.

Johansson, I., and Ingelman-Sundberg, M. (2008). CNVs of human genes and their implication in pharmacogenetics. Cytogenet. Genome Res. *123*, 195–204.

Johansson, I., Lundqvist, E., Bertilsson, L., Dahl, M.L., Sjöqvist, F., and Ingelman-Sundberg, M. (1993). Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. Proc. Natl. Acad. Sci. U. S. A. *90*, 11825–11829.

Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T.D., Gong, Z., Gao, H., Hao, K., Willard, M.D., Xu, J., et al. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res. *23*, 1422–1433.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kim, I.-W., Han, N., Kim, M.G., Kim, T., and Oh, J.M. (2015). Copy number variability analysis of pharmacogenes in patients with lymphoma, leukemia, hepatocellular, and lung carcinoma using The Cancer Genome Atlas data. Pharmacogenet. Genomics *25*, 1–7.

Kinirons, M.T., and O'Mahony, M.S. (2004). Drug metabolism and ageing. Br. J. Clin. Pharmacol. *57*, 540–544.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. *40*, e69–e69.

Klein, K., Winter, S., Turpeinen, M., Schwab, M., and Zanger, U.M. (2010). Pathway-Targeted Pharmacogenomics of CYP1A2 in Human Liver. Front. Pharmacol. *1*.

Klein, M., Thomas, M., Hofmann, U., Seehofer, D., Damm, G., and Zanger, U.M. (2014). A Systematic Comparison of the Impact of Inflammatory Signaling on ADME Gene Expression and Activity in Primary Human Hepatocytes and HepaRG Cells. Drug Metab. Dispos. Biol. Fate Chem.

Koenker, R. (2015). quantreg: Quantile Regression. R package version 5.11. http://CRAN.R-project.org/package=quantreg.

Koukouritaki, S.B., Manro, J.R., Marsh, S.A., Stevens, J.C., Rettie, A.E., McCarver, D.G., and Hines, R.N. (2004). Developmental expression of human hepatic CYP2C9 and CYP2C19. J. Pharmacol. Exp. Ther. *308*, 965–974.

Kramer, W.E., Walker, D.L., O'Kane, D.J., Mrazek, D.A., Fisher, P.K., Dukek, B.A., Bruflat, J.K., and Black, J.L. (2009). CYP2D6: novel genomic structures and alleles. Pharmacogenet. Genomics *19*, 813–822.

Krone, N., Braun, A., Roscher, A.A., Knorr, D., and Schwarz, H.P. (2000). Predicting phenotype in steroid 21-hydroxylase deficiency? Comprehensive genotyping in 155 unrelated, well defined patients from southern Germany. J. Clin. Endocrinol. Metab. *85*, 1059–1065.

Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–1532.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol *9*, e1003118.

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., LaFramboise, T., Brown, M., Tyekucheva, S., and Freedman, M.L. (2013). Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. Cell *152*, 633–641.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu. Rev. Genomics Hum. Genet. *10*, 387–406.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.

Machado, L.R., and Ottolini, B. (2015). An Evolutionary History of Defensins: A Role for Copy Number Variation in Maximizing Host Innate and Adaptive Immune Responses. Front. Immunol. *6*.

Mackenzie, P.I., Bock, K.W., Burchell, B., Guillemette, C., Ikushiro, S., Iyanagi, T., Miners, J.O., Owens, I.S., and Nebert, D.W. (2005). Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. Pharmacogenet. Genomics *15*, 677–685.

Madadi, P., Koren, G., Cairns, J., Chitayat, D., Gaedigk, A., Leeder, J.S., Teitelbaum, R., Karaskov, T., and Aleksa, K. (2007). Safety of codeine during breastfeeding: fatal morphine poisoning in the breastfed neonate of a mother prescribed codeine. Can. Fam. Physician Médecin Fam. Can. *53*, 33–35.

Mandelker, D., Amr, S.S., Pugh, T., Gowrisankar, S., Shakhbatyan, R., Duffy, E., Bowser, M., Harrison, B., Lafferty, K., Mahanta, L., et al. (2014). Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. J. Mol. Diagn. JMD *16*, 639–647.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.

Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R.J., and Scott, S.A. (2013). Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. Pharmacogenomics J. *13*, 558–566.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. (2006). Common deletion polymorphisms in the human genome. Nat. Genet. *38*, 86–92.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.

McGuire, M.M., Bowden, W., Engel, N.J., Ahn, H.W., Kovanci, E., and Rajkovic, A. (2011). Genomic analysis using high-resolution single-nucleotide polymorphism arrays reveals novel microdeletions associated with premature ovarian failure. Fertil. Steril. *95*, 1595–1600.

McLellan, R.A., Oscarson, M., Alexandrie, A.K., Seidegård, J., Evans, D.A., Rannug, A., and Ingelman-Sundberg, M. (1997). Characterization of a human glutathione S-transferase mu cluster containing a duplicated GSTM1 gene that causes ultrarapid enzyme activity. Mol. Pharmacol. *52*, 958–965.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods *6*, S13–S20.

Ménard, V., Eap, O., Harvey, M., Guillemette, C., and Lévesque, É. (2009). Copy-number variations (CNVs) of the human sex steroid metabolizing genes UGT2B17 and UGT2B28 and their associations with a UGT2B15 functional polymorphism. Hum. Mutat. *30*, 1310–1319.

Meyer, U.A. (1996). Overview of enzymes of drug metabolism. J. Pharmacokinet. Biopharm. *24*, 449–459.

Mullis, K.B., and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. *155*, 335–350.

Mwenifumbo, J.C., and Tyndale, R.F. (2007). Genetic variability in CYP2A6 and the pharmacokinetics of nicotine. Pharmacogenomics *8*, 1385–1402.

Mwenifumbo, J.C., Lessov-Schlaggar, C.N., Zhou, Q., Krasnow, R.E., Swan, G.E., Benowitz, N.L., and Tyndale, R.F. (2008). Identification of novel CYP2A6*1B variants: the CYP2A6*1B allele is associated with faster in vivo nicotine metabolism. Clin. Pharmacol. Ther. *83*, 115–121.

Nakano, M., Mohri, T., Fukami, T., Takamiya, M., Aoki, Y., McLeod, H.L., and Nakajima, M. (2015). SNPs in CYP2E1 3'-UTR affect the regulation of CYP2E1 by miR-570. Drug Metab. Dispos. Biol. Fate Chem.

National Toxicology Program (2000). NTP Toxicology and Carcinogenesis Studies of Methyleugenol (CAS NO. 93-15-2) in F344/N Rats and B6C3F1 Mice (Gavage Studies). Natl. Toxicol. Program Tech. Rep. Ser. *491*, 1–412.

Neafsey, P., Ginsberg, G., Hattis, D., Johns, D.O., Guyton, K.Z., and Sonawane, B. (2009). Genetic polymorphism in CYP2E1: Population distribution of CYP2E1 activity. J. Toxicol. Environ. Health B Crit. Rev. *12*, 362–388.

Nelson, D.R., Zeldin, D.C., Hoffman, S.M.G., Maltais, L.J., Wain, H.M., and Nebert, D.W. (2004). Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics *14*, 1–18.

Novak, R.F., and Woodcroft, K.J. (2000). The alcohol-inducible form of cytochrome P450 (CYP 2E1): role in toxicology and regulation of expression. Arch. Pharm. Res. *23*, 267–282.

Ohno, S., and Nakajin, S. (2011). Quantitative analysis of UGT2B28 mRNA expression by real-time RT-PCR and application to human tissue distribution study. Drug Metab. Lett. *5*, 202–208.

Ohtsuki, S., Schaefer, O., Kawakami, H., Inoue, T., Liehner, S., Saito, A., Ishiguro, N., Kishimoto, W., Ludwig-Schwellinger, E., Ebner, T., et al. (2012). Simultaneous absolute protein quantification of transporters, cytochromes P450, and UDP-glucuronosyltransferases as a novel approach for the characterization of individual human liver: comparison with mRNA levels and activities. Drug Metab. Dispos. Biol. Fate Chem. *40*, 83–92.

Oneta, C.M., Lieber, C.S., Li, J., Rüttimann, S., Schmid, B., Lattmann, J., Rosman, A.S., and Seitz, H.K. (2002). Dynamics of cytochrome P4502E1 activity in man: induction by ethanol and disappearance during withdrawal phase. J. Hepatol. *36*, 47–52.

Parajes, S., Quinteiro, C., Domínguez, F., and Loidi, L. (2008). High frequency of copy number variations and sequence variants at CYP21A2 locus: implication for the genetic diagnosis of 21-hydroxylase deficiency. PloS One *3*, e2138.

Pellé, L., Cipollini, M., Tremmel, R., Romei, C., Figlioli, G., Gemignani, F., Melaiu, O., De Santi, C., Barone, E., Elisei, R., et al. (2016). Association between CYP2E1 polymorphisms and risk of differentiated thyroid carcinoma. Arch. Toxicol.

Piehler, A.P., Hellum, M., Wenzel, J.J., Kaminski, E., Haug, K.B.F., Kierulf, P., and Kaminski, W.E. (2008). The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. BMC Genomics *9*, 165.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R.F. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? RNA N. Y. N *17*, 792–798.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Raftogianis, R.B., Wood, T.C., and Weinshilboum, R.M. (1999). Human phenol sulfotransferases SULT1A2 and SULT1A1: genetic polymorphisms, allozyme properties, and human liver genotype-phenotype correlations. Biochem. Pharmacol. *58*, 605–616.

Raimundo, S., Fischer, J., Eichelbaum, M., Griese, E.U., Schwab, M., and Zanger, U.M. (2000). Elucidation of the genetic basis of the common “intermediate metabolizer” phenotype for drug oxidation by CYP2D6. Pharmacogenetics *10*, 577–581.

Rao, Y., Hoffmann, E., Zia, M., Bodin, L., Zeman, M., Sellers, E.M., and Tyndale, R.F. (2000). Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol. Pharmacol. *58*, 747–755.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Riches, Z., Stanley, E.L., Bloomer, J.C., and Coughtrie, M.W.H. (2009). Quantitative Evaluation of the Expression and Activity of Five Major Sulfotransferases (SULTs) in Human Tissues: The SULT “Pie.” Drug Metab. Dispos. *37*, 2255–2261.

Riedmaier, S., Klein, K., Hofmann, U., Keskitalo, J.E., Neuvonen, P.J., Schwab, M., Niemi, M., and Zanger, U.M. (2010). UDP-glucuronosyltransferase (UGT) polymorphisms affect atorvastatin lactonization in vitro and in vivo. Clin. Pharmacol. Ther. *87*, 65–73.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140.

Rose-Zerilli, M.J., Barton, S.J., Henderson, A.J., Shaheen, S.O., and Holloway, J.W. (2009). Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR. Clin. Chem. *55*, 1680–1685.

Rouprêt, M., Cancel-Tassin, G., Comperat, E., Fromont, G., Sibony, M., Molinié, V., Allory, Y., Triau, S., Champigneulle, J., Gaffory, C., et al. (2007). Phenol sulfotransferase SULT1A1*2 allele and enhanced risk of upper urinary tract urothelial cell carcinoma. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. *16*, 2500–2503.

Sadee, W., Wang, D., Papp, A.C., Pinsonneault, J.K., Smith, R.M., Moyer, R.A., and Johnson, A.D. (2011). Pharmacogenomics of the RNA world: structural RNA polymorphisms in drug therapy. Clin. Pharmacol. Ther. *89*, 355–365.

Sanghani, S.P., Sanghani, P.C., Schiel, M.A., and Bosron, W.F. (2009). Human carboxylesterases: an update on CES1, CES2 and CES3. Protein Pept. Lett. *16*, 1207–1214.

Sathirapongsasuti, J.F., Lee, H., Horst, B.A.J., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J., and Nelson, S.F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics *27*, 2648–2654.

Schaeffeler, E., Schwab, M., Eichelbaum, M., and Zanger, U.M. (2003). CYP2D6 genotyping strategy based on gene copy number determination by TaqMan real-time PCR. Hum. Mutat. *22*, 476–485.

Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. Nat. Genet. *39*, S7–S15.

Schinkel, A.H., and Jonker, J.W. (2003). Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. Adv. Drug Deliv. Rev. *55*, 3–29.

Schröder, A., Klein, K., Winter, S., Schwab, M., Bonin, M., Zell, A., and Zanger, U.M. (2013). Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. Pharmacogenomics J. *13*, 12–20.

Schulze, J.J., Lundmark, J., Garle, M., Skilving, I., Ekström, L., and Rane, A. (2008). Doping test results dependent on genotype of uridine diphospho-glucuronosyl transferase 2B17, the major enzyme for testosterone glucuronidation. J. Clin. Endocrinol. Metab. *93*, 2500–2506.

Schuster-Böckler, B., Conrad, D., and Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. PloS One *5*, e9474.

Seidegård, J., and Ekström, G. (1997). The role of human glutathione transferases and epoxide hydrolases in the metabolism of xenobiotics. Environ. Health Perspect. *105*, 791–799.

Settels, E., Bernauer, U., Palavinskas, R., Klaffke, H.S., Gundert-Remy, U., and Appel, K.E. (2008). Human CYP2E1 mediates the formation of glycidamide from acrylamide. Arch. Toxicol. *82*, 717–727.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.

Shibata, T., and Aburatani, H. (2014). Exploration of liver cancer genomes. Nat. Rev. Gastroenterol. Hepatol. *11*, 340–349.

Sprenger, R., Schlagenhaufer, R., Kerb, R., Bruhn, C., Brockmöller, J., Roots, I., and Brinkmann, U. (2000). Characterization of the glutathione S-transferase GSTT1 deletion: discrimination of all genotypes by polymerase chain reaction indicates a trimodular genotype-phenotype correlation. Pharmacogenetics *10*, 557–565.

Staudinger, J.L., and Lichti, K. (2008). Cell signaling and nuclear receptors: new opportunities for molecular pharmaceuticals in liver disease. Mol. Pharm. *5*, 17–34.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

Strott, C.A. (2002). Sulfonation and molecular action. Endocr. Rev. *23*, 703–732.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE *6*, e21800.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Toscano, C., Raimundo, S., Klein, K., Eichelbaum, M., Schwab, M., and Zanger, U.M. (2006). A silent mutation (2939G>A, exon 6; CYP2D6*59) leading to impaired expression and function of CYP2D6. Pharmacogenet. Genomics *16*, 767–770.

Tourancheau, A., Margaillan, G., Rouleau, M., Gilbert, I., Villeneuve, L., Lévesque, E., Droit, A., and Guillemette, C. (2015). Unravelling the transcriptomic landscape of the major phase II UDP-glucuronosyltransferase drug metabolizing pathway using targeted RNA sequencing. Pharmacogenomics J.

Tremmel, R., Klein, K., Winter, S., Schaeffeler, E., and Zanger, U.M. (2015). Gene copy number variation analysis reveals dosage-insensitive expression of CYP2E1. Pharmacogenomics J.

Turner, S.D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv 005165.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419.

Veitia, R.A., Bottani, S., and Birchler, J.A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet. TIG *24*, 390–397.

Vieira, I., Sonnier, M., and Cresteil, T. (1996). Developmental Expression of CYP2E1 in the Human Liver. Eur. J. Biochem. *238*, 476–483.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

Wang, K., Lim, H.Y., Shi, S., Lee, J., Deng, S., Xie, T., Zhu, Z., Wang, Y., Pocalyko, D., Yang, W.J., et al. (2013). Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. Hepatology *58*, 706–717.

Ward, L.D., and Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. gkr917.

Webb, A., Lind, P.A., Kalmijn, J., Feiler, H.S., Smith, T.L., Schuckit, M.A., and Wilhelmsen, K. (2011). The investigation into CYP2E1 in relation to the level of response to alcohol through a combination of linkage and association analysis. Alcohol. Clin. Exp. Res. *35*, 10–18.

Wells, P.G., Mackenzie, P.I., Chowdhury, J.R., Guillemette, C., Gregory, P.A., Ishii, Y., Hansen, A.J., Kessler, F.K., Kim, P.M., Chowdhury, N.R., et al. (2004). Glucuronidation and the Udp-Glucuronosyltransferases in Health and Disease. Drug Metab. Dispos. *32*, 281–290.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

White, P.C., and Speiser, P.W. (2000). Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Endocr. Rev. *21*, 245–291.

Wolbold, R., Klein, K., Burk, O., Nüssler, A.K., Neuhaus, P., Eichelbaum, M., Schwab, M., and Zanger, U.M. (2003). Sex is a major determinant of CYP3A4 expression in human liver. Hepatol. Baltim. Md *38*, 978–988.

Woodwark, C., and Bateman, A. (2011). The characterisation of three types of genes that overlie copy number variable regions. PloS One *6*, e14814.

Xu, S., Wang, Y., Roe, B., and Pearson, W.R. (1998). Characterization of the human class Mu glutathione S-transferase gene cluster and the GSTM1 deletion. J. Biol. Chem. *273*, 3517–3527.

Xue, Y., Sun, D., Daly, A., Yang, F., Zhou, X., Zhao, M., Huang, N., Zerjal, T., Lee, C., Carter, N.P., et al. (2008). Adaptive evolution of UGT2B17 copy-number variation. Am. J. Hum. Genet. *83*, 337–346.

Yang, D., Pearce, R.E., Wang, X., Gaedigk, R., Wan, Y.-J.Y., and Yan, B. (2009). Human Carboxylesterases HCE1 and HCE2: Ontogenic Expression, Inter-Individual Variability and Differential Hydrolysis of Oseltamivir, Aspirin, Deltamethrin and Permethrin. Biochem. Pharmacol. *77*, 238–247.

Yang, T.-L., Guo, Y., Shen, H., Li, J., Glessner, J.T., Qiu, C., Deng, F.-Y., Tian, Q., Yu, P., Liu, Y.-Z., et al. (2013). Copy Number Variation on Chromosome 10q26.3 for Obesity Identified by a Genome-Wide Study. J. Clin. Endocrinol. Metab. *98*, E191–E195.

Yu, X., Dhakal, I.B., Beggs, M., Edavana, V.K., Williams, S., Zhang, X., Mercer, K., Ning, B., Lang, N.P., Kadlubar, F.F., et al. (2010). Functional genetic variants in the 3'-untranslated region of sulfotransferase isoform 1A1 (SULT1A1) and their effect on enzymatic activity. Toxicol. Sci. Off. J. Soc. Toxicol. *118*, 391–403.

Zanger, U.M. (2012). Introduction to Drug Metabolism. In Metabolism of Drugs and Other Xenobitics, (Weinheim, Germany: Wiley-VCH), pp. 287–300.

Zanger, U.M., and Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol. Ther. *138*, 103–141.

Zanger, U.M., Fischer, J., Raimundo, S., Stüven, T., Evert, B.O., Schwab, M., and Eichelbaum, M. (2001). Comprehensive analysis of the genetic factors determining expression and function of hepatic CYP2D6. Pharmacogenetics *11*, 573–585.

Zanger, U.M., Klein, K., Richter, T., Toscano, C., and Zukunft, J. (2005). Impact of genetic polymorphism in relation to other factors on expression and function of human drug-metabolizing p450s. Toxicol. Mech. Methods *15*, 121–124.

Zanger, U.M., Turpeinen, M., Klein, K., and Schwab, M. (2008). Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. Anal. Bioanal. Chem. *392*, 1093–1108.

Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics *14 Suppl 11*, S1.

Zhou, J., Lemos, B., Dopman, E.B., and Hartl, D.L. (2011). Copy-Number Variation: The Balance between Gene Dosage and Expression in Drosophila melanogaster. Genome Biol. Evol. *3*, 1014–1024.

Publikationen

Zanger, U.M., Klein, K., Thomas, M., Rieger, J.K., **Tremmel, R**., Kandel, B.A., Klein, M., and Magdy, T. (2014). Genetics, epigenetics, and regulation of drug-metabolizing cytochrome p450 enzymes. Clin. Pharmacol. Ther. 95, 258–261.

Tremmel, R., Klein, K., Winter, S., Schaeffeler, E., and Zanger, U.M. (2015). Gene copy number variation analysis reveals dosage-insensitive expression of CYP2E1. Pharmacogenomics J. [Epub ahead of print]

Pellé, L., Cipollini, M., **Tremmel, R.**, Romei, C., Figlioli, G., Gemignani, F., Melaiu, O., De Santi, C., Barone, E., Elisei, R., Seiser, E., Innocenti, F., Zanger, U.M., Landi, S. (2016). Association between CYP2E1 polymorphisms and risk of differentiated thyroid carcinoma. Arch. Toxicol. [Epub ahead of print]

Wissenschaftliche Beiträge

- **R Tremmel**, K Klein, M Schwab, UM. Zanger; Functional assessment of CNV's in human liver; Posterpräsentation auf dem 17. North American Regional ISSX Meeting 2011, Atlanta, USA
- **R Tremmel**, K Klein, S Winter, M Schwab, UM. Zanger; CYP2E1 in human liver: a gene-dosage insensitive gene; Posterpräsentation auf dem 10. International ISSX Meeting 2013, Toronto, Kanada
- S Fehr, F Battke, T Scheurenbrand, K Klein, **R Tremmel,** E Schaeffeler, M Schwab, UM Zanger, S Biskup; Highly efficient screening of ADME genes using panel based next-generation sequencing; Poster auf dem 20. International Symposium on Microsomes and Drug Oxidations 2014, Stuttgart
- **R Tremmel,** S Fehr, K Klein, E Schaeffeler, M Schwab, S Biskup, UM Zanger; ADME-wide Gene Copy Number Variations and their Functional Relevance in Human Liver; Posterpräsentation auf dem 20. International Symposium on Microsomes and Drug Oxidations 2014, Stuttgart
- K Klein, S Fehr, **R Tremmel**, E Schaeffeler , S Winter, M. Schwab, S. Biskup, UM. Zanger; Targeted exome resequencing of ADME genes in human liver: assessment of SNV/CNV frequencies and possible functional relevance for cytochrome P450's; Poster auf dem 81. Annual congress of the german society for experimental and clinical pharmacology and toxicology DGPT 2015, Kiel
- **R Tremmel**, S Fehr, F Battke, K Klein, E Schaeffeler, M Schwab, S Biskup, UM. Zanger; ADME-wide analysis of copy number variation using targeted exome resequencing and their functional relevance in human liver; Posterpräsentation auf dem 12. Congress of the European Association for Clinical Pharmacology and Therapeutics EACPT 2015, Madrid, Spanien
- K Klein, S Fehr, **R Tremmel**, E Schaeffeler , S Winter, M. Schwab, S. Biskup, UM. Zanger; Targeted exome resequencing: ADME Pharmacogenetics in human liver; Poster auf dem 12. Congress of the European Association for Clinical Pharmacology and Therapeutics EACPT 2015, Madrid, Spanien
- H Glatt, W Meinl, W Engst, F Schumacher, B Sachse, K Herrmann, G Barknowitz, M Bernau, C Bendadani, M Wiesner, M Schreiner, **R Tremmel**, A Bub, U Zager, B Monien; Natural and process-related carcinogens in food: Macromolecular adducts in animal models and human blood and tissue samples. Poster auf dem 51st Congress of the European Societies of Toxicology (EUROTOX), Porto, Portugal

Danksagung

Ohne der Hilfe, die Geduld und Unterstützung folgender Personen wäre diese Arbeit nicht vollendet worden.

Professor Dr. Ulrich M. Zanger danke ich für die Möglichkeit in diesem interessanten Forschungsbereich zu arbeiten. Außerdem bedanke ich mich für die hilfreichen wissenschaftliche Diskussionen, die lektorischen Überarbeitungen, aber auch persönliche Ratschläge während meiner Promotion. So konnte ich neue Techniken lernen und in das weite Feld der Bioinformatik eintauchen, was die tägliche Arbeit spannend und aufregend machte.

Mein Dank gilt auch Professor Dr. Matthias Schwab, dem Leiter des Dr. Margarete Fischer-Bosch Instituts für klinische Pharmakologie für die Möglichkeit, meine Dissertation an diesem Institut durchzuführen und in dieser tollen Umgebung weiter Forschung betreiben zu können.

Bei Professor Dr. Lutz Graeve bedanke ich mich herzlich für die Bereitschaft die Betreuung und Begutachtung meiner Dissertation von Seiten der Universität Hohenheim zu übernehmen.

Bei Dr. Kathrin Klein bedanke ich mich besonders für Ihre Hilfsbereitschaft und die Unterstützung in wissenschaftlichen und methodischen Fragen als auch für viele Tipps, Anregungen und kreative Impulse für Vorträge, Veröffentlichungen und Kongresse sowie für tolle Stuttgart-Tipps.

Ich danke der Arbeitsgruppe für die schöne Zeit. Dabei gilt besonderer Dank Britta Klumpp und Igor Liebermann für die technische Unterstützung und Hilfsbereitschaft. Bei Dr. Jessica Rieger, Dr. Marcus Klein und Dr. Benjamin Kandel möchte ich mich für die gute Zusammenarbeit, für die fachlichen und persönlichen Gespräche und die gemeinsamen Besuche von Kongressen und Urlauben bedanken. Bei Dr. Maria Thomas möchte ich mich für Ihren wissenschaftlichen Input, für Diskussionen und ihre Unterstützung bedanken.

Dr. Stefan Winter und Dr. Florian Büttner danke ich für die Hilfeleistung und die Geduld bei Fragen zur Bioinformatik und Statistik.

Bedanken möchte ich mich auch sehr herzlich bei den Kooperationspartnern Prof. Dr. Stefano Landi und Lucia Pellè von der Universität Pisa in Italien und Prof. Dr. Hans-Rudolf Glatt, sowie Dr. Walter Meinl vom Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke (DIfE) in Nuthetal.

Ein großer Dank geht an die Freunde und Kollegen am IKP. Insbesondere der Fußballgruppe, die mir durch den sportlichen Ausgleich und durch viele Diskussionen über den VfB den Alltag verschönerte.

Meinen Freunden danke ich für die Unterstützung und den Zusammenhalt in allen Lebenslagen.

Meiner Partnerin Sabine Hieronymus möchte ich für Ihre Geduld, die Liebe, Freude und Motivation während der letzten Jahre danken.

Zum Schluss danke ich ganz besonders meinen Eltern und meinen Geschwistern für Ihre Unterstützung und Hilfe während meiner Schulzeit, meines Studiums und der Promotion. Ohne diesen Rückhalt wäre diese Arbeit nicht möglich gewesen.

Eidesstattliche Versicherung

Eidesstattliche Versicherung gemäß § 7 Absatz 7 der Promotionsordnung der Universität Hohenheim zum Dr. rer. nat.

1. Bei der eingereichten Dissertation zum Thema

Bioinformatische Analyse und Funktionelle Charakterisierung von Strukturellen Genvarianten in ADME-Genen in Humaner Leber

handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

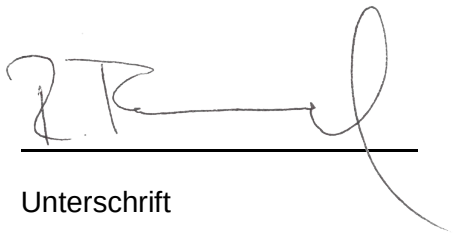
3. Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.

4. Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Die Richtigkeit der vorstehenden Erklärung bestätige ich: Ich versichere an Eides Statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Stuttgart, den 26.02.2016

Ort, Datum



Unterschrift

Lebenslauf

	PERSÖNLICH
Name	**Roman Tremmel**
Geburtsdatum	20.02.1984 / Ostfildern-Ruit
Familienstand	Ledig
	BERUFSERFAHRUNG
seit 2015	**Wissenschaftlicher Angestellter**
	Dr. Margarete Fischer-Bosch-Institut für
	Klinische Pharmakologie
	AUSBILDUNG
seit 2011	**PROMOTION** **Bioinformatische Analyse und Funktionelle Charakterisierung von Strukturellen Genvarianten in ADME-Genen in Humaner Leber**
	Dr. Margarete Fischer-Bosch-Institut für Klinische Pharmakologie
2005/2010	**STUDIUM DIPLOM BIOLOGIE**
	Komplementärmutation in *Notch* und *Suppressor of Hairless* zur Analyse exogener Su(H)-Aktivität
	Universität Hohenheim
	Hauptfach: Genetik
	Nebenfächer: Zoologie & Tierphysiologie
2004/2005	**ZIVILDIENST**
	Klinikum Esslingen
1995/2004	**GYMNASIUM**
	Schelzorgymnasium Esslingen
	Abitur

Anhang

Tabelle S1: ADME-Genloci, die von mindestens zwei CNVs komplett eingeschlossen waren (DGV Version 16.10.2014).

Gen	CNV-Anzahl[1]	Durchschnittliche Frequenz [%][2]		Referenzen[3]	ADME-Gruppe[4]
		Duplikation	Deletion		
CYP2E1	30	5,21	1,07	1, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14, 17, 18	Phase I
GSTT1	19	6,04	22,20	1, 3, 4, 6, 7, 8, 12, 14, 16, 17, 19	Phase II
UGT2B28	18	3,18	22,05	1, 3, 4, 6, 7, 8, 11, 12, 16, 17, 19	Phase II
UGT2B17	16	5,79	18,90	1, 3, 4, 6, 7, 8, 12, 19	Phase II
GSTT2B	13	8,01	32,60	1, 3, 6, 7, 12, 14, 16, 17, 19	Phase II
CYP4F12	12	3,15	0,22	3, 6, 7, 9, 10, 13, 14, 17, 18	Phase I
SULT1A1	12	1,95	2,50	6, 7, 8, 16, 19	Phase II
GSTA1	11	0,99	17,56	1, 7, 9, 10, 13, 16, 17, 19	Phase II
CYP21A2	10	3,53	5,55	6, 7, 16	Phase I
CYP2A6	9	8,28	10,71	1, 6, 7, 8, 14, 17, 18, 19	Phase I
GSTM1	9	12,23	29,56	1, 3, 4, 6, 7, 12, 16, 17, 19	Phase II
UGT2B11	8	6,60	30,97	1, 6, 7, 16, 19	Phase II
AHR	7	0,21	0,03	7, 9, 13, 15, 17	Modifizierer
ALDH3B1	6	0,01	0,18	7, 9, 10	Phase I
GSTA2	6	1,78	0,05	7, 9, 10, 13, 16	Phase II
NUDT8	6	0,11	0,02	7, 9, 15	ADME-Verwandte
TYMS	6	0,01	0,01	7, 14	ADME-Verwandte
ABCC6	5	8,74	29,84	7, 8, 9, 16, 19	Transporter
ABCF1	5	0,11	20,02	1, 2, 7, 19	Transporter
ALDH2	5	0,12	NA	7, 9, 10, 14, 17	Phase I
CYP2A7	5	14,66	5,92	1, 6, 7, 8, 17	Phase I
SULT1C2	5	0,24	12,34	1, 13, 14, 19	Phase II
UGT2B7	5	2,06	19,18	7, 14, 16, 19	Phase II
ABCA11P	4	0,02	0,02	7, 9	Transporter
ABCA2	4	-	1,56	7, 10, 14, 15	Transporter
CYP2C18	4	-	8,37	1, 7, 14, 19	Phase I
CYP4F2	4	0,14	12,24	1, 2, 19	Phase I
GPS2	4	-	0,20	1, 7, 9, 15	Modifizierer
KCNJ11	4	0,00	0,29	7, 9, 14	Modifizierer

Gen	CNV-Anzahl[1]	Durchschnittliche Frequenz [%][2]		Referenzen[3]	ADME-Gruppe[4]
		Duplikation	Deletion		
MIF	4	0,02	23,70	7, 14, 19	ADME-Verwandte
SLC19A1	4	0,02	0,13	7, 14, 17	Transporter
SLC22A12	4	0,22	0,16	7, 14, 15, 18	Transporter
SLC22A18AS	4	-	0,25	7, 9, 10	Transporter
SLC2A4	4	-	0,20	1, 7, 9, 15	Transporter
STK19	4	0,00	0,09	7	ADME-Verwandte
TRPM7	4	0,15	0,00	7, 9, 10	Transporter
UGT2B15	4	25,70	4,43	6, 7, 16, 19	Phase II
ABCA3	3	0,18	0,02	7, 9, 15	Transporter
ABCC1	3	0,03	28,82	7, 9, 19	Transporter
ABCC13	3	0,18	NA	7, 9, 10	Transporter
ALDH3A1	3	-	0,25	7, 9, 10	Phase I
ALDH3B2	3	0,21	NA	7, 9, 15	Phase I
ALDH8A1	3	0,03	NA	7, 9	Phase I
CABIN1	3	0,02	31,60	7, 14, 19	ADME-Verwandte
CYP19A1	3	0,30	NA	7, 10, 13	Phase I
CYP27B1	3	0,08	0,08	7, 10	Phase I
CYP46A1	3	0,00	0,08	7, 10	Phase I
GPX4	3	0,23	0,08	7, 10	Phase I
GSTA5	3	3,44	3,20	10, 16, 19	Phase II
SLCO1B1	3	0,02	0,14	7, 9, 14	Transporter
SLCO4A1	3	0,21	NA	7, 9, 15	Transporter
SULT2A1	3	0,02	0,00	7, 9	Phase II
UGT2B4	3	0,10	NA	7, 9, 10	Phase II
ABCA5	2	-	0,01	7	Transporter
ABCB1	2	0,04	NA	7, 9	Transporter
ABCB4	2	0,04	NA	7, 9	Transporter
ABCC9	2	0,12	NA	7, 10	Transporter
ABCD1	2	5,44	38,51	6, 8	Transporter
ABCG4	2	-	0,04	7, 9	Transporter
ABCG5	2	0,04	NA	7, 9	Transporter
ABCG8	2	0,04	NA	7, 9	Transporter
AHRR	2	0,04	NA	7, 9	Modifizierer
BSG	2	0,00	1,02	7, 10	ADME-Verwandte
CYP11B2	2	-	0,16	1, 10	Phase I
CYP2B6	2	4,64	31,77	16, 19	Phase I

Gen	CNV-Anzahl[1]	Durchschnittliche Frequenz [%][2]		Referenzen[3]	ADME-Gruppe[4]
		Duplikation	Deletion		
CYP2B7P1	2	-	33,44	5, 19	Phase I
CYP2C19	2	-	16,68	7, 19	Phase I
CYP2D6	2	8,75	51,04	6, 19	Phase I
CYP3A7	2	5,16	5,21	16, 19	Phase I
CYP4A11	2	5,20	NA	16, 17	Phase I
CYP4F3	2	-	40,41	1, 19	Phase I
EAF2	2	0,03	1,15	8, 14	ADME-Verwandte
ESRRA	2	-	0,12	7, 10	Modifizierer
GSTM2	2	35,00	12,50	06:07	Phase II
GSTP1	2	0,28	NA	7, 15	Phase II
NCOR1	2	0,03	34,38	14, 19	Modifizierer
PRMT1	2	0,56	NA	10, 18	ADME-Verwandte
RXRA	2	-	50,45	10, 19	Modifizierer
SLC10A1	2	0,04	NA	7, 9	Transporter
SLC15A2	2	0,03	1,15	8, 14	Transporter
SLC22A1	2	0,77	NA	13, 18	Transporter
SLC22A13	2	-	28,34	1, 19	Transporter
SLC22A2	2	0,77	NA	13, 18	Transporter
SLC22A25	2	-	1,57	7, 19	Transporter
SLC22A3	2	0,77	NA	13, 18	Transporter
SLC22A7	2	-	0,01	7	Transporter
SLC29A4	2	5,16	0,01	7, 16	Transporter
SLC2A5	2	-	0,04	7, 9	Transporter
SLC7A5	2	0,04	NA	7, 9	Transporter
VKORC1	2	-	0,04	7, 9	ADME-Verwandte

[1]Anzahl an CNVs die die Genregion komplett einschließen (Positionen hg19). [2]Angegeben ist die populationsunabhängige mittlere Frequenz. [3]Die Nummern beziehen sich auf die Studien in Tabelle S2. [4]Gruppeneinteilung nach ADME-Gen Funktion.

Tabelle S2: Verwendete DGV-Arbeiten zur Analyse von ADME-CNVs. In Fett markiert sind Studien die für die CNV-Analyse unter anderen HapMap-Proben verwendet haben.

Studie	PubmedID	Methode	Probenanzahl
1) 1000 Genomes Consortium Phase 1	**23128226**	**Oligo-Mikrochip, PCR, Sequenzierung, SNP-Mikrochip**	**1151**
2) 1000 Genomes Consortium Pilot Project	**20981092**	**Digitaler Mikrochip, Oligo-Mikrochip, PCR, Sequenzierung**	**185**

Studie	PubmedID	Methode	Probenanzahl
3) Altshuler et al. 2010	**20811451**	**SNP-Mikrochip**	**1184**
4) Campbell et al. 2011	**21397061**	**Oligo-Mikrochip**	**2366**
5) Conrad et al. 2006	**16327808**	**Oligo-Mikrochip, SNP-Mikrochip**	**60**
6) Conrad et al. 2009	**19812545**	**Oligo-Mikrochip**	**40**
7) Cooper et al. 2011	21841781	Oligo-Mikrochip, SNP-Mikrochip	17421
8) de Smith et al. 2007	17666407	Oligo-Mikrochip	51
9) Itsara et al. 2009	19166990	Oligo-Mikrochip, SNP-Mikrochip	1557
10) Jakobsson et al. 2008	18288195	SNP-Mikrochip	443
11) McCarroll et al. 2006	**16468122**	**SNP-Mikrochip**	**269**
12) McCarroll et al. 2008	**18776908**	**SNP-Mikrochip**	**270**
13) Pinto et al. 2007	**17911159**	**SNP-Mikrochip**	**771**
14) Shaikh et al. 2009	19592680	SNP-Mikrochip	2026
15) Simon-Sanchez et al. 2007	17116639	qPCR, SNP-Mikrochip	181
16) Sudmant et al. 2013	23825009	Oligo-Mikrochip, Sequenzierung	97
17) Vogler et al. 2010	21179565	SNP-Mikrochip	1109
18) Wang et al. 2007	**17921354**	**SNP-Mikrochip**	**112**
19) Wong et al. 2012b	23290073	Sequenzierung	96
20) Banerjee et al. 2011[a]	21479260	SNP-Mikrochip	1250
21) Chia et al. 2012[a]	23635498	Karyotypisierung, PCR, Sequenzierung, SNP-Mikrochip	64

[a]Studien deren detektierten CNVs keine der 340 ADME-Gene einschließen.

Tabelle S3: Paarweiser Sequenzvergleich der *CYP2D6* und *CYP2D7* Exons

Exon	Waterman-Eggert Wert	bits	E(1)	Similarität
E1& P	5608	393.9	4.20E-113	94.0%
E1	1384	167.1	4.40E-46	96.9%
E2	788	99.5	5.40E-26	95.3%
E3	720	101.6	6.00E-27	96.7%
E4	787	123.7	1.50E-33	98.8%
E5	840	140.3	1.80E-38	97.2%
E6	647	112.2	3.40E-30	95.1%
E7	908	151.2	1.30E-41	98.4%

E8	701	135.6	3.10E-37	99.3%
E9	1128	137.5	2.50E-37	94.9%

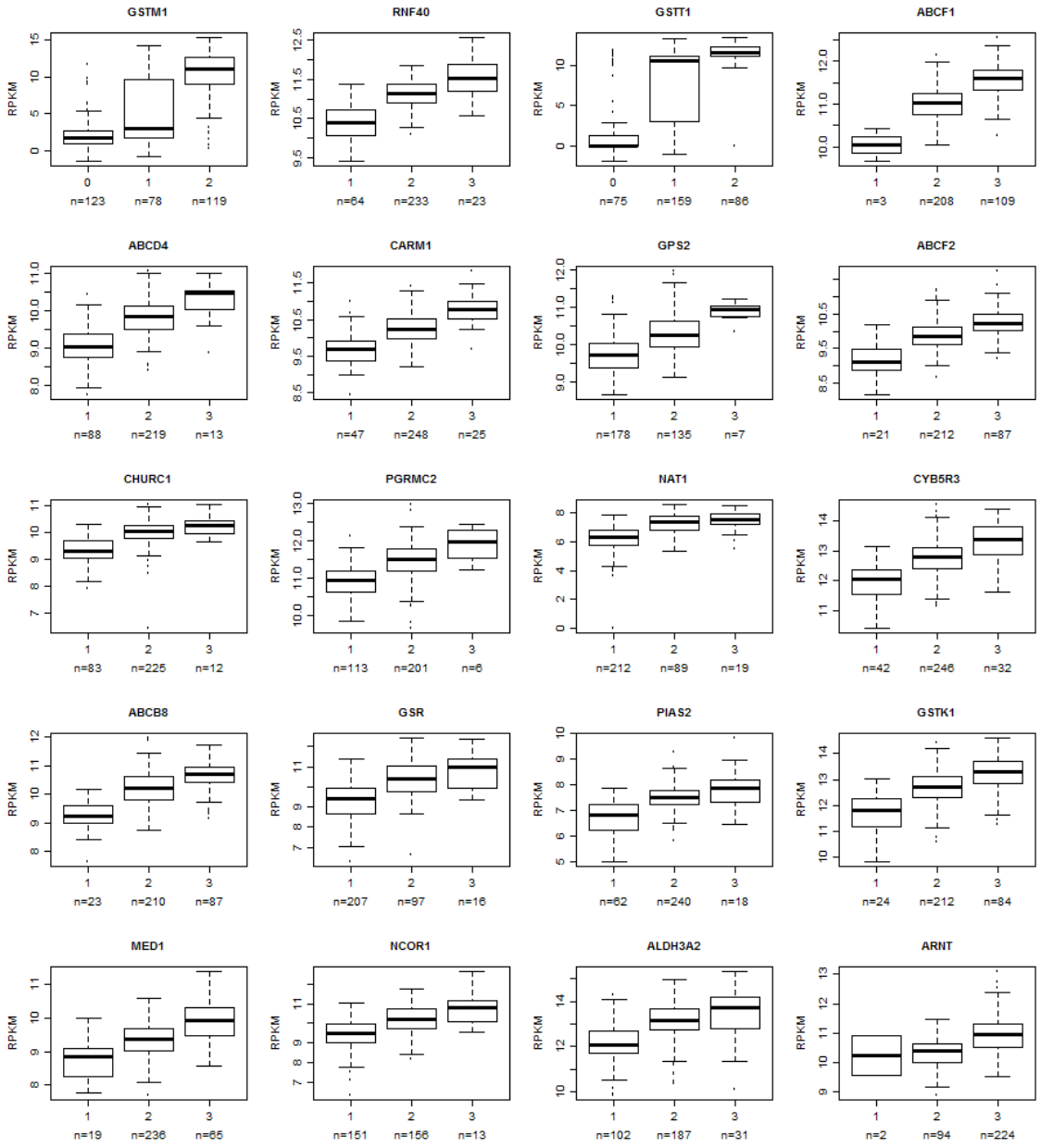


Abbildung A1: Boxplotdiagramme der Top 20 zur mRNA Expression (RNA Sequenzierung) assoziierten CNVs, die mit dem Affymetrix 6.0 Mikrochip in TCGA-Tumorgewebe bestimmt wurden (n=343).

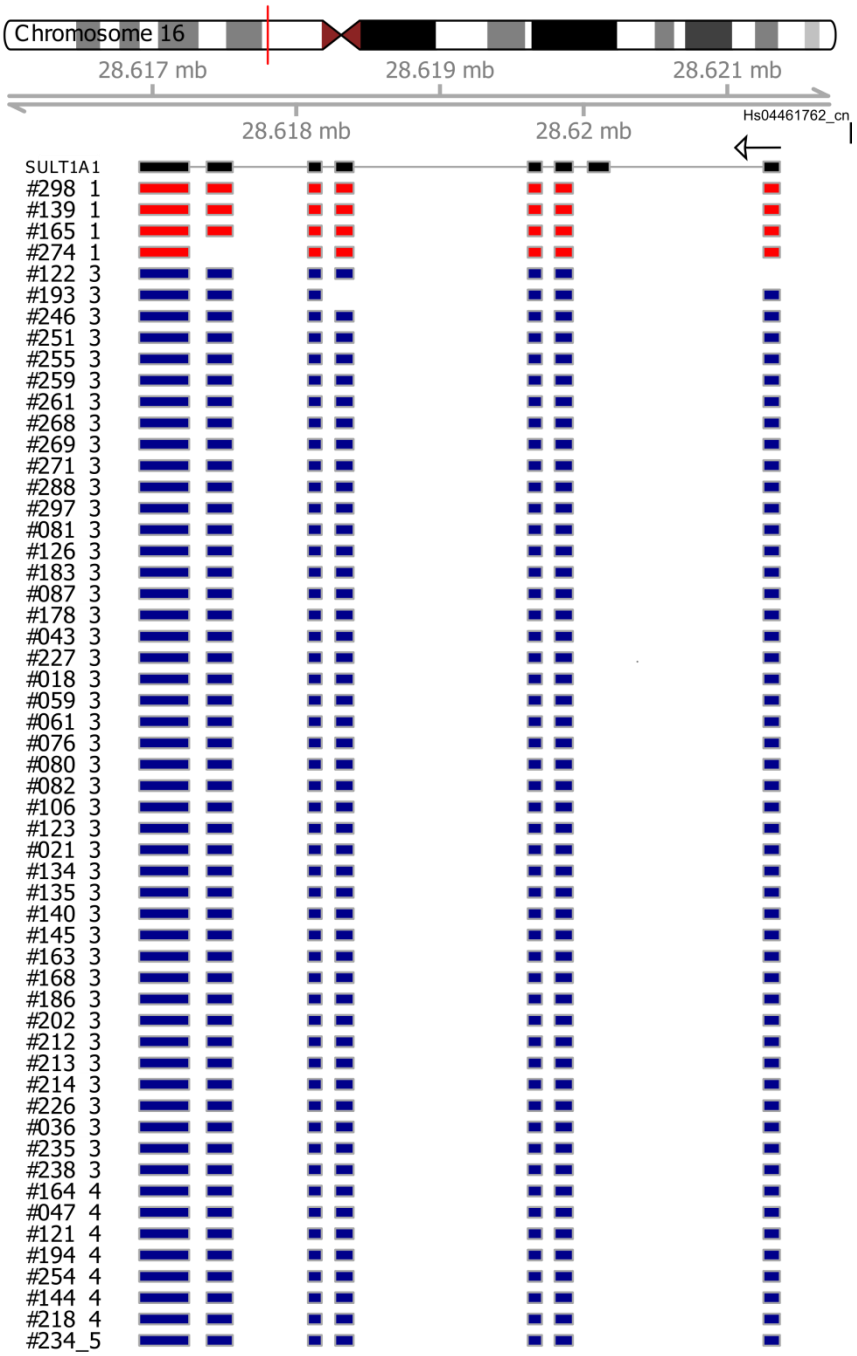


Abbildung A2: Von oben nach unten sind ein Ideogramm des Chromosoms 16, die Genomposition (GRCh37, hg19: Feb. 2009), die einzelnen Exons des Gens *SULT1A1* als schwarze Boxen, der TaqMan Assay und die Ergebnisse der Abdeckungsanalyse auf Exonebene. In Rot sind deletierte Sequenzen markiert. In blau sind duplizierte Exons markiert. Die bestimmte Kopienzahl ist hinter der Probenkennzeichnung angegeben.