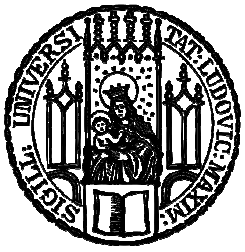
GENETIC AND FUNCTIONAL CHARACTERIZATION OF CANDIDATE GENES FOR COMPLEX PSYCHIATRIC DISEASES USING NEXT-GENERATION SEQUENCING AND CELLULAR UPTAKE ASSAYS

Dissertation der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

zur Erlangung des Doktorgrades der
Naturwissenschaften Dr. rer. nat.

angefertigt am Max-Planck-Institut für Psychiatrie München



vorgelegt von
Carina Quast
10. Juli 2013

Erstgutachter: Prof. Dr. Rainer Landgraf

Zweitgutachter: Prof. Dr. Wolfgang Enard

Tag der mündlichen Prüfung: 22. Januar 2014

TABLE OF CONTENTS

ABSTRACT

Complex phenotypes are the result of a complex interplay between genes and environmental factors. Extensive linkage, candidate and genome-wide association studies (GWASs) have been carried out to unravel genetic risk variants for human diseases. The identification of genes, involved in the pathomechanism of a disease, might be beneficial for its diagnosis, treatment and prognosis. While GWASs allowed the identification of a large number of common variants robustly associated with common complex diseases [1,2], the heritability, which can be explained by these variants, is small [3]. The discrepancy between the estimated heritability from twin, family and adoption studies and the heritability obtained from GWAS was termed “missing heritability” and led to the investigation of additional factors that might also contribute to disease susceptibility, including gene-environment interactions, gene-gene interactions, structural variants and rare variants.

In this thesis, the role of less common and rare variants in susceptibility to common complex diseases was investigated. In order to accomplish this, a candidate gene for panic disorder (PD) and a possible risk gene for major depressive disorder (MDD) were screened for the presence of common and rare variants using next-generation sequencing in a pooled approach. In a previously published GWAS, a haplotype containing two common intronic variants in the transmembrane protein 132D (*TMEM132D*) gene was associated with PD [4]. Another GWAS identified solute carrier family 6 member 15 (*SLC6A15*), which encodes an amino acid transporter, as a risk gene for MDD [5]. A common intergenic variant about 600 kilobase downstream of this gene was shown to decrease *SLC6A15* gene expression in lymphoblastoid cell lines and hippocampus. Susceptibility genes for complex diseases, identified in GWAS, are promising candidates for the search of rare variants as genes harbouring common variants are likely to contain also rare variants [6].

Pooled targeted re-sequencing of the exonic regions of *TMEM132D* in 300 anxiety disorder patients, mostly suffering from PD (84.7%), and 300 healthy controls allowed the detection of 371 genetic variants. Of these variants, 24.0% were common (minor allele frequency (MAF) > 5.0%), whereas the vast majority was less common (MAF 1.0 – 5.0%) to rare. 247 variants had not been reported before, including 12 novel non-synonymous variants leading to an amino acid exchange in the protein. While common variants associated with PD were not identified, an overrepresentation of non-synonymous variants and variants with predicted changes on splicing in healthy controls

compared to PD patients was observed. These putatively functional relevant variants were distributed along a broad MAF spectrum, ranging from 0.17 to 30.0%. In addition, a higher rate of private non-synonymous variants, which were only present in either cases or controls in this study, but not in over 7,500 individuals with different ethnic backgrounds from other publicly available re-sequencing datasets, in patients compared to controls was seen. Combined with the data from the previous GWAS study in which the association with PD was carried by common variants [4], this pooled re-sequencing study suggests that not only common or rare variants alone, but a combination of both contributes to the development of anxiety-related phenotypes.

Re-sequencing the whole *SLC6A15* locus in 400 MDD patients and 400 healthy controls, 405 genetic variants were identified, including twelve non-synonymous variants. Only 15.0% of the detected variants were common. While none of the non-synonymous variants was significantly associated with MDD, two rare non-synonymous variants were identified to influence protein function. In contrast to the TMEM132D protein whose molecular function has still to be discovered, SLC6A15 is known to transport neutral amino acids into predominantly neuronal cells [7]. The cellular uptake of neutral amino acids such as proline is thus a measurable property that associates with function. The uptake experiments identified two rare variants to be associated with a significant increase in proline uptake in HEK cells. This result suggests that rare variants in *SLC6A15* might influence the biochemical function of its amino acid transporter and thus downstream neuronal function and possibly the risk for MDD and other stress-related psychiatric disorders. In addition, this study highlights that functional exploration of genetic variants might be promising to identify putatively disease-relevant variants as statistically significant associations for rare variants might only be achieved in extremely large samples.

ZUSAMMENFASSUNG

Komplexe Phänotypen sind das Ergebnis eines komplexen Zusammenspiels von Genen und Umweltfaktoren. Umfangreiche Linkage-, Kandidatengen- und genomweite Assoziationsstudien wurden durchgeführt um genetische Risikovarianten für humane Erkrankungen zu entdecken. Die Identifizierung von Genen, die am Pathomechanismus einer Erkrankung beteiligt sind, könnte förderlich sein für deren Diagnose, Behandlung und Prognose. Obwohl genomweite Assoziationsstudien viele häufige Varianten identifiziert haben, die mit häufigen komplexen Erkrankungen assoziiert sind, ist die Erblichkeit, die durch diese Varianten erklärt werden kann, klein. Die Diskrepanz zwischen der Erblichkeit, die in Zwillings-, Familien- und Adoptionsstudien geschätzt wurde, und der Erblichkeit, die in genomweiten Assoziationsstudien ermittelt wurde, wurde als „fehlende Erblichkeit“ bezeichntet. Diese „fehlende Erblichkeit“ führte zu der Untersuchung von zusätzlichen Faktoren wie seltenen Varianten, strukturellen Varianten, Gen-Umwelt Interaktionen und Gen-Gen Interaktionen, die ebenfalls zu der Suszeptibilität für eine Erkrankung beitragen könnten.

In dieser Arbeit wurde die Rolle von seltenen Varianten in der Suszeptibilität für häufige komplexe Erkrankungen untersucht. Dazu wurde unter Verwendung der neusten Sequenziertechnologie (Next-Generation Sequencing) nach häufigen und seltenen Varianten innerhalb eines Kandidatengens für Panikstörung und eines möglichen Risikogens für Depression gesucht. In einer kürzlich veröffentlichten genomweiten Assoziationsstudie war ein Haplotyp, der aus zwei häufigen intronischen Varianten im Transmembranprotein 132D *(TMEM132D)* Gen besteht, mit Panikstörung assoziiert. In einer anderen genomweiten Assoziationsstudie wurde das *SLC6A15* Gen, das für einen Aminosäuretransporter kodiert, als Risikogen für Depression identifiziert. Eine häufige Variante, die ca. 600 Kilobasen von diesem Gen entfernt ist, führte zu einer verringerten SLC6A15 Genexpression in lymphoblastioden Zelllinien und dem Hippocampus. Suszeptibilitätsgene für komplexe Erkrankungen, die in genomweiten Assoziationsstudien identifiziert wurden, sind erfolgversprechende Kandidaten für die Suche nach seltenen Varianten da in Genen, die häufige Varianten tragen, wahrscheinlich auch seltene Varianten vorhanden sind.

Das gezielte Re-Sequenzieren der exonischen Regionen des *TMEM132D* Gens in 300 Angstpatienten, die größtenteils unter Panikstörung litten (84.7%), und 300 gesunden Kontrollen, die jeweils in Pools zusammengefasst wurden, ermöglichte es 371 genetische Varianten zu identifizieren. Von diesen Varianten waren 24.0% häufig

(Frequenz des veränderten Allels > 5.0%), wohingegen der Großteil der Varianten geringere Frequenzen hatte. 247 Varianten wurden zuvor noch nicht berichtet, darunter 12 nicht synonyme Varianten, die zu einem Aminosäureaustausch im Protein führen. Während keine häufigen Varianten identifiziert wurden, die mit Panikstörung assoziiert sind, konnte in gesunden Kontrollen eine Überrepräsentierung von nicht synonymen Varianten und Varianten mit möglichen Veränderungen auf das Spleißen („Splicing“) im Vergleich zu Panikpatienten beobachtet werden. Die Frequenz der veränderten Allelle dieser möglicherweise funktionell relevanten Varianten war zwischen 0.17 und 30.0%. Zusätzlich konnte eine höhere Rate von „private“ nicht synonymen Varianten in Patienten im Vergleich zu Kontrollen beobachtet werden. „Private“ Varianten waren ausschließlich in Patienten oder Kontrollen dieser Studie und nicht in über 7500 Individuen unterschiedlichster ethnischer Herkunft, die in anderen Re-Sequenzierungsprojekten untersucht wurden, vorhanden. Kombiniert mit den Daten der genomweiten Assoziationsstudie, in der eine Assoziation von häufigen Varianten mit Panikstörung identifiziert wurde, zeigt diese Re-Sequenzierungsstudie, dass nicht nur häufige oder seltene Varianten alleine sondern eine Kombination aus beidem zu der Entstehung von Phänotypen, die mit Angst assoziiert sind, beiträgt.

Das Re-Sequenzieren des gesamten *SLC6A15* Lokus in 400 depressiven Patienten und 400 gesunden Kontrollen identifizierte 405 genetische Varianten einschließlich zwölf nicht synonymer Varianten. Nur 15.0% der detektieren Varianten waren häufig. Während keine der nicht synonymen Varianten mit Depression assoziiert war, wurden zwei seltene nicht synonyme Varianten identifiziert, die die Funktion des Proteins beeinflussen. Im Gegensatz zum TMEM132D Protein, dessen Funktion noch entschlüsselt werden muss, ist für SLC6A15 bekannt, dass es neutrale Aminosäuren in überwiegend neuronale Zellen transportiert. Die zelluläre Aufnahme von neutralen Aminosäuren wie beispielsweise Prolin ist somit eine messbare Größe, die die Funktion des Proteins wiederspiegelt. Die Aufnahmeexperimente identifizierten zwei seltene Varianten, die mit einer signifikant erhöhten Prolinaufnahme in HEK-Zellen assoziiert waren. Dieses Ergebnis zeigt, dass seltene Varianten in *SLC6A15* die biochemische Funktion des Aminosäuretransporters und somit die neuronale Funktion und möglicherweise das Risiko für Depression und andere stressbezogene psychiatrische Erkrankungen beeinflussen könnten. Zusätzlich zeigt diese Studie, dass die funktionelle Untersuchung von genetischen Varianten erfolgversprechend sein könnte um krankheitsrelevante Varianten zu identifizieren, da signifikante Assoziationen von seltenen Varianten nur in extrem großen Studienkohorten beobachtet werden können.

1. INTRODUCTION

1.1 Genetics of complex diseases

1.1.1 Genotype-phenotype relationship

The aim of genetics is to understand the relationship between genotypes and phenotypes [8]. The extent to which the phenotype of an individual is determined by the genotype is different between traits. Some traits are highly influenced by genetic factors and rather independent of the environment, whereas other traits are more determined by the environment and only to a less extent by the genotype (Figure 1).

Monogenic disorders are the classical examples of traits which are highly determined by the genotype, whereas environmental influences play a minor role. This class of disorders which are often referred to as Mendelian diseases due to their Mendelian pattern of inheritance is assumed to be caused by genetic variants in one single gene (monogenic) irrespective of environmental exposures. Furthermore, the penetrance, which is the proportion of individuals who carry a disease causing variant of a gene and develop the associated phenotype, is expected to be high. Prominent examples for these relatively rare monogenic diseases are Huntington disease [9], cystic fibrosis [10], phenylketonuria [11] and fragile X syndrome [12].

Complex diseases are characterized by a contribution of multiple genes (polygenic) and environmental factors to the phenotype. Mostly, the penetrance is incomplete as the presence of one altered gene alone is not sufficient to develop the associated disease. Therefore, altered genes which are related to complex traits are referred to as susceptibility, vulnerability or risk genes instead of causal genes. The substantial influence of the environment on a phenotype has been demonstrated by Caspi *et al*. among others. In their study, individuals carrying a risk gene for depression showed more depressive symptoms compared to individuals who did not carry the risk gene, but only in the presence of stressful life events [13]. Other examples of highly prevalent complex diseases are diabetes [14,15], cardiovascular disease [15] and psychiatric diseases, including schizophrenia [16] and bipolar disorder [15,17].

For a long time, diseases were subdivided in the two classes described above. Nowadays, it is becoming more and more obvious that the line between Mendelian and complex disease is not always clear. For instance, it is known that the monongenic disease phenylketonuria manifests only in individuals with variants in the disease causing gene that are exposed to a distinct environment, which in this case is phenylalanine in the diet. Furthermore, it has been shown that the severity of the

monogenic diseases cystic fibrosis and sickle cell anemia is modulated by additional genes [18,19]. The scenario in which a single gene causes a complex disorder does also exist. Genetic variants in the *BRCA1* [20] and *BRCA2* [21] gene were reported to cause breast cancer in families, although in most cases breast cancer is a multifactorial disease. These examples show that the phenotype of an individual is always the result of a complex interplay between nature (genotype) and the nurture (environment) [22].

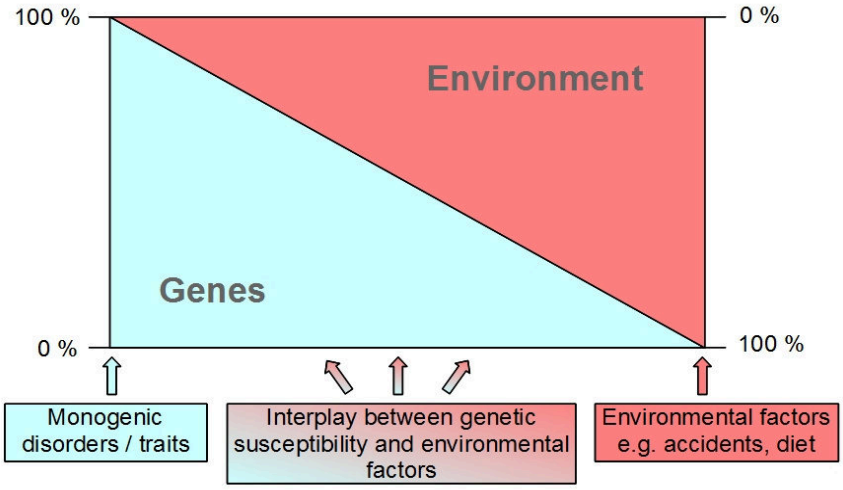


Figure 1 Influence of genotype and environment on an individual's phenotype.
In most cases the phenotype is determined by a complex interplay between both genes and environment to different proportions. Adapted from Murken [23].

1.1.2 Genetic epidemiology

One measure to quantify the contribution of genes to the determination of a phenotype is the heritability. It is the proportion of total variance in a population that can be explained by genetic variation among the individuals in the population [24]. High heritability scores imply a strong correlation between genotype and phenotype, whereas lower scores suggest a higher impact of environmental factors.

The heritability of a disease can be estimated by family, twin and adoption studies [25-27]. In family studies the prevalence of a disorder among relatives of an affected or unaffected family member is examined. Especially siblings or half-siblings are often used as study subjects since they share 50% or 25% of their genes respectively. The ratio between the prevalence of a disease among relatives of an affected family member and the prevalence among relatives of a healthy individual is a measure for the influence of genetic components on the disease, with high ratios indicating high heritability.

Twin pairs are valuable study subjects as they share both genetic and environmental factors. Monozygotic (MZ) twins are derived from one fertilized egg and are thus genetic

identical, whereas dizygotic (DZ) twins share only 50% of their genes since they originate from two different zygotes. In addition, uterus, birth date, age and parts of the early and late environment are shared in both MZ and DZ twin pairs [28]. Heritability is estimated by comparing the concordance rate for a disease in MZ and DZ twins. As MZ twins have the same genetic background they also share the same susceptibility genes which leads to the assumption that the disease manifests in both twins. Thus, a higher concordance rate in MZ twins compared to DZ twins indicates that the disease is highly influenced by genetic factors. MZ twins that are discordant for a disease offer the opportunity to estimate the degree to which non-genetic effects determine the phenotype.

Adoption studies are another method to estimate heritability. Children separated from their biological parents and raised in the home of their adoptive parents have an environmental, but not a genetic change so that the genetic similarity of the children to their biological as well as to their adoptive parents can be investigated. If similar risks for a disorder can be observed in adopted children and their biological parents, a genetic influence is suggested, whereas similar risks for adoptees and adoptive parents indicate that the shared environment might be the main disease causing factor.

1.1.3 Classes of human variation in the genome

The human genome consists of 2.85 billion nucleotides and harbours 20,000 – 25,000 genes [29]. About 1% of the nucleotides are located in exons and 24% in introns, whereas the vast majority of the genome consists of intergenic DNA [30]. In 2001, the first human reference genomes were published. First, an assembly of sequences from different donors was released by the Human Genome Sequencing Consortium [31]. Second, Celera Genomics published a consensus sequence derived from five individuals [30]. In 2007, the first complete genome of an individual, Craig Venter, was released [32]. Individual sequences can thus be mapped to a reference genome which allows the identification of nucleotides that differ among individuals.

Genetic variants are usually subdivided into common and rare, depended on the frequency of the minor allele in the human population. Common variants, which are also referred to as polymorphisms, are variants with a minor allele frequency (MAF) of at least 1% in the general population, whereas rare variants have a MAF less than 1% [33]. Furthermore, genetic variants can be subdivided into single nucleotide variants (SNVs) and structural variants, based on the class of variation [34]. SNVs are DNA sequence variants in which a single nucleotide is affected. SNVs within coding regions of the

genome might, but do not necessarily have to, have effects on the function of the gene product. The genetic code is redundant which means that an amino acid is encoded by several codons. A nucleotide substitution which does not change the amino acid is called synonymous or silent variant. In contrast, non-synonymous variants might have an influence on protein function as they either lead to an amino acid exchange which is referred to as missense variant, or a premature stop codon which is called nonsense variant.

SNVs are the most prevalent class of genetic variation among individuals. Recently, the 1000 Genomes Project sequenced the genomes of about 1,100 individuals from 14 different populations [35]. This sequencing effort allowed the identification of 38 million SNVs in the human population and 3.7 million per individual. Thus, on average, there is about one SNV every 75 base pairs (bp) throughout the human genome, while in an individual genome every 770 bp a SNV is present. Information about SNVs can be gained from the Single Nucleotide Polymorphism Database (dbSNP) which catalogues all variants throughout the genome of different species that have been submitted, regardless of their frequency and functional consequences [36].

The second class of genetic variation is structural variation which was originally defined to affect more than 1 kb of DNA sequence [37]. Structural variants include insertions, deletions, inversions, translocations and copy number variants (CNVs) as described in Table 1.

Table 1 Structural variation in the human genome.

Class of structural variation	Definition
Insertion	Incorporation of extra bases into the DNA
Deletion	Removal of DNA bases from the genome
Inversion	Chromosomal break at two places of the DNA, incorporation of the reversed DNA segment into the same chromosome
Translocation	Exchange of DNA segments between different chromosomes
CNV	Repeat of identical DNA sequences

In the past, the insight into location, frequency and functional consequences of structural variants was rather low due to the lack of adequate detection methods. Recently developed massive parallel sequencing methods [38,39] and complement microarray-based methods [40] allowed the discovery of a large number of structural variants, even smaller variations with a length of > 50 bp, leading to a novel definition for structural variants [40].

This class of variants contributes to a larger extent to differences in the human genome than SNVs. Although only 20% of all genetic variants are structural variants, these account for more than 70% of the altered nucleotides, implicating an important role of structural variants in human health and diseases [33,41-43].

1.1.4 Approaches for mapping genetic variants

To identify genetic variants that either increase the risk for a disease or protect against it, two prerequisites have to be fulfilled: first, the disease of interest has to be heritable. For complex psychiatric disorders, the estimated heritability ranges from moderate to high, with heritability scores of 0.28 for panic disorder (PD) [44], 0.37 for depression [45], 0.67 for schizophrenia [46] and 0.75 for attention deficit hyperactivity disorder (ADHD) [47]. Second, a dense set of genetic variants spanning the whole genome is necessary. As the genome harbours one SNV per every 75 bp, this class of genetic variation is commonly used for genetic dissection of diseases. The two pivotal methods to accomplish this are linkage and association analyses.

1.1.4.1 Linkage analysis

Linkage is defined as “the existence or establishment of connection of two things” [48]. In the context of genetics, linkage is the co-segregation of a genetic marker and a disease which requires that the genetic variant responsible for the disease is located in the region where the marker is located (Figure 2).

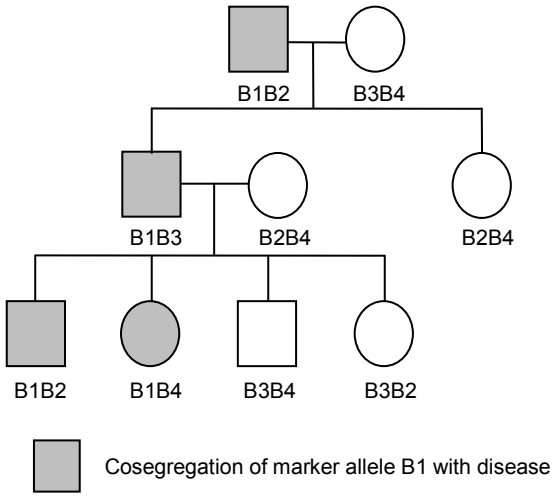


Figure 2 Linkage study design.

Marker allele B1 cosegregates with disease in a family consisting of three generations. The squares indicate males and the circles females. The grey squares and circles denote affected individuals. Adapted from Kullo and Ding [49].

The prerequisite for this is that the order of the genetic markers on a chromosome is known. In 1910, Morgan observed that the recombination rate (crossover) between two markers in gametes differs leading to the idea that the frequency of crossover events indicates the distance between markers on the chromosome [50]. Sturtevant confirmed that the greater the crossover rate the greater the distance between two loci as the probability that these loci are separated by an exchange of genetic material between two homologue chromosomes is increased [51,52]. This idea built the basis for the first linkage map. As recombination events occur in gametes, the effects become first visible in the next generation. Thus, linkage analyses can only be carried out in families and requires information about the disease status in all family members.

Genome-wide linkage studies usually use 300 – 600 genetic markers, with marker to marker distances of 10 and 5 centimorgan (cM) respectively [53]. cM is defined as the distance between two loci where the expected number of crossover events in a single generation is 0.01. The markers of choice for linkage analyses are simple tandem repeats (STRs) which are also referred to as variable number tandem repeats (VNTRs). STRs are di-, tri- or tetranucleotide repeats widely and evenly distributed across the genome [54]. The variable number of repeats indicates high mutation rates [55] making them so useful for studying co-segregation of marker and disease.

Linkage analyses have been traditionally performed to map genes responsible for monogenic disorders. Unfortunately, linkage studies have only sufficient power to detect genes with large effects on the phenotype. As complex disorders are comprised of multiple genes with small effect sizes, the recruitment of an extremely large sample would be necessary to reach adequate statistical power [56]. Nevertheless, linkage studies can also play a role in mapping susceptibility genes for complex disorders. The Mendelian subtypes of complex diseases described in section 1.1.1 are caused by one single gene with a large effect which can be detected much easier than multiple genes with subtle effects. Although such subtypes are rare, several linkage studies were successful in identifying disease-relevant genes, for instance the Alzheimer's disease relevant genes beta-amyloid precursor protein [57] and presenilin 1 and 2 [58,59]. Such findings could give new insights into the pathomechanism of the Mendelian subtype and as implication also into the more common forms of the disease.

1.1.4.2 Association studies

The term association derives from the medieval Latin word “associare” which means “to connect”. Genetic association is the simultaneous occurrence of a genetic variant and a trait. If an association is present then an allele or a genotype of a variant will be seen more often than expected by chance in individuals carrying the trait. Association differs from linkage in that the frequency of a genetic marker is compared between unrelated affected individuals and unaffected controls, while linkage is based on the investigation of co-inheritence of chromosomal regions with a phenotype in families [60]. Although the commonly used approach to test for association is a case-control design in which unrelated individuals are investigated, association testing can also be performed in families. Association studies have greater power than linkage studies so that genes with small effects can be detected, however, a much higher number of markers has to be examined. Computer simulations have been estimated that 500,000 markers are necessary for a genome-wide analysis [60]. The higher number of required markers can be explained by the fact that in population-based studies the order of the genetic markers on the chromosomes varies as a result of many recombination events over a high number of generations. In linkage studies, only three or four generations of pedigrees are examined and thus much less recombination occurs resulting in a relative stable marker map [54].

For association analyses the markers of choice are single nucleotide polymorphisms (SNPs). This class of genetic variation is very frequent in the human genome and more stable than STRs due to a lower mutation rate [61]. Furthermore, SNPs might have functional consequences if they are located in coding or regulatory regions of the genome, whereas variations in STRs rarely contribute to the trait [54].

Genetic association can be subdivided into two classes: direct and indirect association. Direct association means that the genotyped SNP itself is the causal variant contributing to disease susceptibility (Figure 3a). Testing variants for direct association is indicated when information about possible consequences of the associated variant on gene function is available. Variants leading to amino acid exchanges or truncated proteins are the most reasonable variants for direct association analyses. However, only about 1% of the human genome consists of protein-coding sequences and coding variants account for only 20% of the associated variants, the vast majority falls outside coding regions [1]. Non-coding variants may influence gene regulation [61-63], differential splicing [64,65] or gene expression [66,67]. So far, the evaluation of intronic and intergenic variants is a particular challenge as the non-coding genome is poorly annotated [68]. Thus, direct

association testing might only discover a small proportion of genetic variants associated with disease.

The second class of association is indirect association, which describes the scenario, in which the genotyped markers are not themselves involved in disease risk, but variants in close proximity to the genotyped ones (Figure 3b). In this context, association is based on the principle of linkage disequilibrium (LD) which is the non-random association of alleles at two or more loci [69]. Generally, loci that are physically close together have a stronger LD than loci that are far away on a chromosome which means that the stronger the LD the higher the association between the two loci. If a marker and a causal variant are in LD and the sample size is adequate, significant association can be detected by only genotyping the marker, although directly genotyping the causal variant should have a higher power to unravel associations [70].

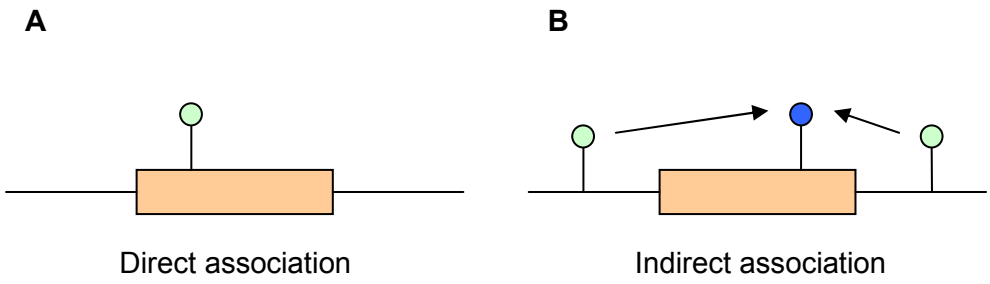


Figure 3 Direct and indirect genetic association.

A A genotyped SNP (green) in a candidate gene (orange box) is directly tested for association with a disease phenotype. This strategy is indicated when prior knowledge about possible functional consequences of a variant is available. B The blue SNP is tested for association indirectly as it is in LD with the two genotyped SNPs (green). Adapted from Hirschhorn and Daly [71].

Indirect association testing has the advantage that *a priori* candidate variants are not required as the decision which marker to genotype is based on information about LD structure. Several studies revealed that the human genome is structured into discrete LD blocks [72,73]. Regions with high LD are separated by smaller regions containing hot spots of recombination which breakdown the LD. Markers for indirect association studies are choosen in that way that all LD blocks in the region of interest are sufficiently covered by the so called tagging SNPs. The International HapMap Project sequenced individuals from four geographically different human populations in order to characterize the genome-wide LD structure and thus to identify an optimized panel of tagging SNPs for association studies by avoiding redundant SNPs [74]. The HapMap Consortium showed that 1.09 million SNPs are required to capture all common SNPs with LD

strength $r^2 \geq 0.8$ in Africans ($r^2 = 1$ perfect LD), whereas 500,000 tagging SNPs are sufficient in Europeans and Asians which is consistent with the number calculated by computer simulations [60].

For genetic association studies two approaches exist: a candidate gene approach and a genome-wide approach. Candidate gene approaches are hypothesis-driven in which genes are examined which have already been suggested to be involved in the pathophysiology of a disease, obtained from blood, biopsy or post-mortem gene expression studies in humans or behavioural animal models. This approach has been proven to be extremely powerful for gene-disease search and biomarker and drug target selection [75]. Nevertheless, candidate gene studies are limited by their reliability on existing information about the known or presumed biology of the investigated phenotype and unfortunately, the molecular mechanisms of most biological traits are unknown so far [76]. For instance, if clear information about proteins involved in a specific psychiatric disorder are missing, all genes expressed in the human brain, which is estimated at tens of thousands, are possible candidate genes and have to be investigated [53].

Genome-wide association approaches have the advantage that they are hypothesis-free and thus prior knowledge about candidate genes is not required. The first smaller genome-wide association studies (GWASs) were reported in 2005 [77] and 2006 [78] The first large and well-designed GWAS for complex disorders was performed by the Wellcome Trust Case Control Consortium (WTCCC) in 2007 [15]. Since then, over 1,500 GWAS have been published (see the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies) and thousands of common SNPs robustly associated with common complex traits have been identified. Examples include 71 detected susceptibility loci for Chron's disease [79], 18 vulnerability loci for type 2 diabetes [80] and 40 loci associated with height [81].

1.1.5 Missing heritability and possible explanations

GWASs are one of the most popular study approaches since the last five years and have emerged to be powerful for investigating the genetic architecture of complex diseases [71,82]. Although thousands of common variants associated with common traits and disorders have been turned up and brought light into their genetic complexity [1], most of the variants have small effect sizes and the proportion of heritability explained by these variants is small [3,83]. One prominent example, which illustrates this issue, is human height. This complex trait can be explained by the height of the parents of an individual and thus, the heritability is estimated to be in the range of 80 – 90% [81]. Three large-

scale GWASs identified more than 40 variants associated with differences in height [84-86]. Unfortunately, the associated loci have tiny effects and explain only about 5% of the heritability [81]. Currently, the proportion of heritability that can be explained by common variants is less than 10% for almost all complex traits [33]. Now the question arises, where the missing variants are that underlie these heritable traits. The difference in heritability between estimates from twin, family and adoption studies and estimates derived from association studies is also termed missing heritability and is currently one of the big topics in the genetics of common complex diseases [3,83,87]. Several explanatory models have been suggested and the most important ones are reviewed in the following sections.

1.1.5.1 Rare variants

GWASs are based on the assumption that complex traits can largely be explained by common variants with small to modest effect sizes which is also referred to as common disease-common variant (CDCV) hypothesis [88-90]. Due to the case of missing heritability, the focus has been shifted to the possible role of rare variants in disease susceptibility. The common disease-rare variant (CDRV) hypothesis postulates that multiple rare variants with higher effect sizes contribute to the susceptibility to common complex diseases [91-93].

Traditionally, variants with a MAF below 1% are declared to be rare, but the definition in the literature is arbitrary and varies across studies. In this thesis, rare variants are denoted according to the original definition with 1% as cut-off, variants with a MAF ranging from 1 - 5% are defined as low-frequency or less common variants and common variants are expected to be in 5% of the general population. These rare and low-frequency variants are not sufficiently frequent to be captured by commercial available GWAS genotyping arrays [90,94]. New high-throughput sequencing technologies, which are referred to as next-generation sequencing (NGS), can overcome this limitation as genetic variants of the whole frequency spectrum can be identified. Since their development in 2008, NGS techniques play an important role in numerous fields of application, including genomics, transcriptomics and proteomics. Details on NGS are described in section 1.3.

Irrespective of the MAF, rare variants differ from common polymorphisms in several points which can be explained by population processes in the human lineage [90]. Spontaneously occurring *de novo* mutations can be constantly and rapidly introduced in a population. All *de novo* mutations are initially private which means that they only occur

in the individual harbouring the *de novo* variation, but they can be inherited to the following generations. Originally, rare variants can become common if they are not lost by random genetic drift or filtered out by purifying selection against vulnerability variants over many generations. This leads to two assumptions: First, variants reaching high frequencies in the population are likely to be older than rare variants as the variant frequency is proportional to the number of generations. Thus, the human genome harbours only a few common but many rare variants. It is suggested that on average, only about one common variant per 500 bp in the European population exists [95]. In contrast, population expansion in the past few centuries has resulted in the presence of many rare alleles. The number of *de novo* point mutations is estimated to be around 40 per individual indicating the mass of rare variation within a population [96]. The second assumption is that rare variants have more likely negative effects on the phenotype than common variants [97]. Given that a variant has a high effect size leading to a disease this variant would be negatively selected which corresponds to the low frequency in the population (Figure 4). Indeed, rare variants are twice as much often expected to be non-synonymous and thus potential deleterious than more frequent ones [95]. Common variants have survived negative selection and are thus supposed to have only, if at all, small to modest effects on the phenotype.

The question which often arises in this context is, why some variants conferring to disease risk escaped negative selection and some not. Several explanations exist, both for the persistence of common and rare deleterious variants in the human genome. First, variants associated with disease might survive negative selection when they do not alter the evolutionary fitness. Decreased reproductive rate is one of the most important criteria for negative selection to save the survival of the population. One example is a common variant that confers the risk for nicotine dependence which does not influence the reproduction rate [98]. Furthermore, late-onset disorders such as Alzheimer's disease are also not correlated with a reduction in fitness as the onset of such diseases is after the time of reproduction [99]. In contrast, several neuropsychiatric diseases are characterized by decreased reproduction rates, including schizophrenia, autism and anorexia nervosa [100,101]. Second, variants might cause a disease and protect against another one simultaneously. A common variant in the *ApoL1* gene confers to an increased risk to develop chronic kidney disease in African Americans and protects from Trypanosoma brucei rhodesiense infection at the same time [102]. With regard to rare variants, a third explanation for their persistence in the human gene pool might be that since such variants are so new, for instance only a few generations old, they have not

been subjected to negative selection over a long time. The continued high prevalence of autism and schizophrenia, despite the strong negative selection due to reduced fitness, suggests that *de novo* mutations may be maintaining these disorders in the population [101]. Finally, the mutation rates within a population might be so large that purifying selection can not remove all deleterious variants, so that variants with small to modest effects can drift to higher frequencies in the population [89,103]. As already mentioned, the number of *de novo* mutations is estimated to be around 40 per individual [96] so that it is unlikely that all deleterious variants are removed from a gene pool. For the future, it is suggested that the prevalence of disease will further increase as the selection criteria to remove a variant might be relaxed due to the changed life style in modern humans so that deleterious variants might accumulate [104].

Another characteristic of rare variants is that they are likely to be population specific, while common variants are mostly present in all populations [92]. In addition, rare variants are not in LD with common variants [92]. Overall, investigating rare variants might offer novel insights into disease mechanisms of complex traits [105].

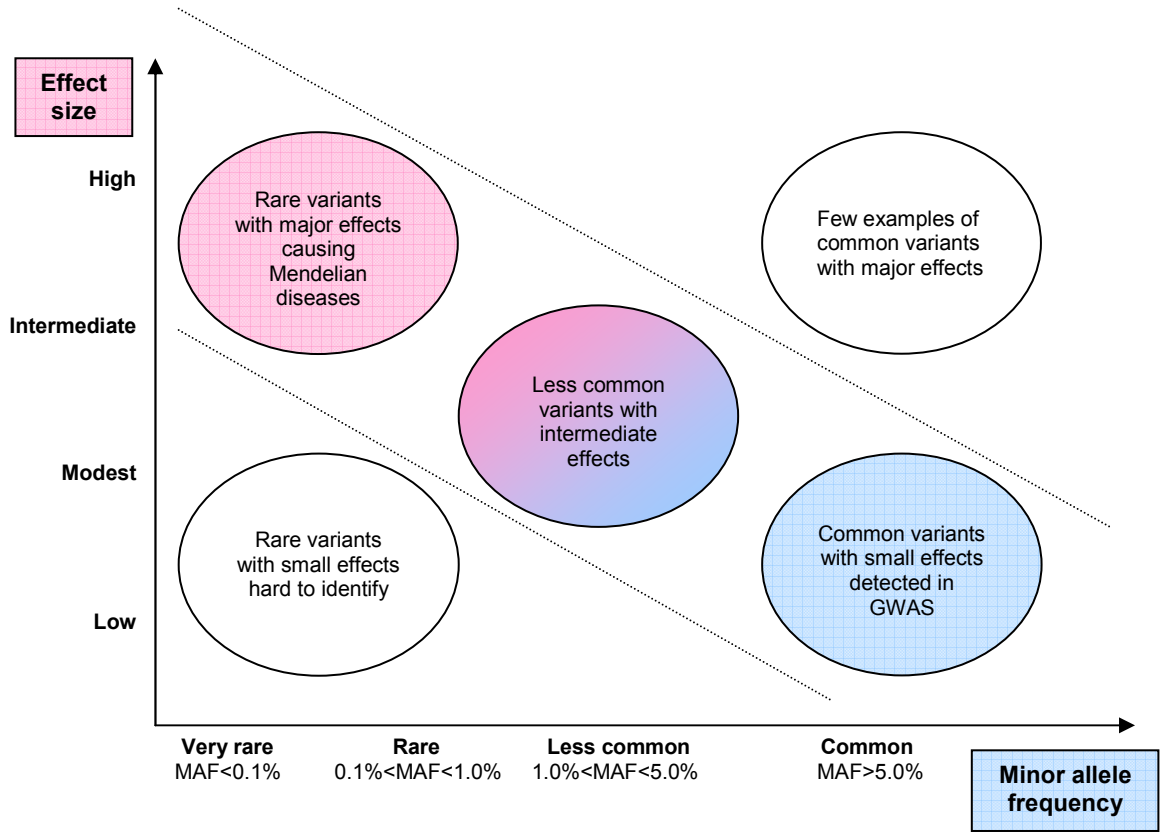


Figure 4 Correlation between allele frequency and effect size.

Effect size and allele frequency are inversely proportional. Rare alleles are assumed to have high effect sizes, while with increased allele frequency the expected effect sizes decrease. Adapted from Manolio *et al.* [83].

1.1.5.2 Structural variants

Although the human genome is rich on structural variants, the discovery and genotyping of this type of genetic variation is far behind those of SNVs [106,107]. Until now, structural variants have not been explicitly examined in most GWASs, although re-sequencing studies have shown that CNVs with a length of at least 1 kb are responsible for the largest variation of the genome between individuals [32]. Variation due to CNVs has been successfully demonstrated to have a functional impact on levels of gene expression [108] and several complex clinical phenotypes. A common 20 kb deletion upstream of the *IRGM* gene was identified to be associated with Crohn's disease [109]. Additional associations could be observed between a common 45 kb deletion upstream of the *NEGR1* gene and body mass index [110] and between a common 32 kb deletion removing the *LCE3B* and *LCE3C* genes and psoriasis [111]. This data lead to the assumption that structural variation might contribute to a considerable proportion of phenotypic variation and thus account for some of the unexplained heritability.

Genetic variation due to CNVs is characterized by a combination of common and rare variants. Similar to SNVs, the majority of variants is rare, but most of the differences between two individuals can be explained by a limited number of common copy number polymorphisms (CNPs) with MAFs ≥ 5% [107]. These CNPs have modest effects on the phenotype and small sizes, ranging from 20 to 45 kb in the above mentioned diseases [83]. In contrast, rare CNVs have higher effect sizes and are much larger than CNPs, ranging from 600 kb to 3 Mb so that they affect many genes [83]. Recently, two studies investigated the distribution of several hundreds of CNVs in healthy individuals and patients suffering from schizophrenia. In these studies, a 1.6 Mb rare deletion with an effect size of 12 associated with disease could be identified [112,113]. Another study identified a rare 600 kb deletion with an OR of 100 and a 600 kb duplication with an OR of 16, both associated with autism [114].

With the increased interest in structural variants as possible explanation for the missing heritability, several approaches have been developed for integrating CNV analysis into GWAS. This includes the design of appropriate genotyping arrays and the use of LD between SNPs and common CNPs. The latter point was initially highly controversial as it was not clear if structural variants are in LD with SNPs at all. However, several studies have shown that common short insertions or deletions with 1 – 5 bp [115-117] and larger common structural variants in unique regions [43,107] of the genome are in LD with tagging SNPs. If CNPs occur in segmental duplications the identification is complicated as the LD with tagging SNPs is low [118]. Segmental duplications are repetitive DNA

sequences that are more than 5 kb in size and have more than 90% inter-copy sequence identity [119]. Low LD in segmental duplications can be explained by the fact that compared with the rest of the genome a small number of validated SNPs that can serve as tagging SNPs exist in such regions [107]. Interestingly, there is a strong relation between structural variation and segmental duplication. 25 – 50% of all nucleotides in large structural variants are located in segmental duplications, which account for only 5.3% of the genome [42]. One explanation for this phenomenon might be that segmental duplications are associated with high rates of non-allelic homologous recombination (NAHR) which occurs when two DNA sequences are to a high extent identical, making these regions more susceptible for rearrangements in general [120]. Thus, it can be concluded that segmental duplications might have an increased probability to harbour disease-relevant variants due to the increased rate of structural variants in these regions.

1.1.5.3 Gene-gene interactions

It has long been recognized that interactions between genes are an important component of genetic architecture of complex traits. In the last years, identification of gene-gene interactions was mostly driven by the failure to identify or to replicate significant associations between variants and a phenotype obtained from linkage or association studies [121,122]. The interplay of several genes is also postulated to explain, in part, the case of missing heritability.

The concept of epistasis or gene-gene interaction describes a masking effect where a variant at one locus prevents the manifestation of the effect of a variant at another locus [123]. An example for gene-gene interactions, which demonstrates the influence on the phenotypic outcome, derives from a study of Cordell (Table 2) [124].

Table 2 Example for gene-gene interaction.

The hair color in mice is determined by the two loci G and B. G encodes grey hairs while B encodes black hairs. Loci G and B have two possible alleles, G/g and B/b. Mice with any copies of the G allele are grey, independent on the genotype of the B locus. The effect of locus B is masked by that of locus G and it is said that locus G is epistatic to locus B. Black mice can only be observed if at least one copy of the dominant B allele, but no G allele is present. Adapted from Cordell [124].

	Genotype locus G		
Genotype locus B	**g/g**	**g/G**	**G/G**
b/b	white	grey	grey
b/B	black	grey	grey
B/B	black	grey	grey

The investigation of gene-gene-interactions might give valuable insights into the pathomechanism of complex diseases as biomolecular interactions are ubiquitous in an organism, comprising gene regulation, signal transduction, biochemical networks and homeostatic developmental and physiological pathways [121]. Under evolutionary aspects, gene-gene interactions are also very meaningful due to the fact that alleles with negative effects can be tolerated as alleles at other loci prevent the manifestation of these negative effects on the phenotype [125]. An example, which illustrates the tolerance of variants with deleterious effects, comes from a study of Gregersen *et al.* who investigated the genetic interactions underlying multiple sclerosis [126]. To accomplish this, transgenic mice expressing either DR2a or DR2b antigen or both were generated. It was found that mice, which only produce DR2b, were highly susceptible to multiple sclerosis, whereas mice only producing the DR2a antigen did not develop the disease. Interestingly, mice in which both gene loci were present showed a reduced disease susceptibility which leads to the assumption that DR2a modulates the effect of DR2b. A possible model for this epistatic effect might be that DR2b stimulates the production of T-cells which are sensitive to the antigen that induces multiple sclerosis, while the DR2a antigen suppresses the T-cell proliferation or induces cell death of these cells. Under normal conditions, the effects of the two antigens balance each other so that the disease-susceptibility antigen DR2b could survive natural selection as only in the absence of DR2a its negative effect manifests in an increased susceptibility risk.

In the literature, many different definitions for the term “epistasis” exist which can be summarized into three main categories [127]. Compositional epistasis is the most similar definition to the original one, representing the prevention of an effect of one allele by an allele at another locus. Functional epistasis describes the molecular interaction of a protein with another one. These protein-protein interactions can be either between proteins that directly interact with each other or between proteins within the same pathway [128]. This definition of epistasis is solely a functional one without a direct genetic link, although a genetic consequence would be predicted in the case of disrupted interactions between proteins [127]. The last definition is statistical epistasis which is used in a quite different sense compared to its original usage. Statistical epistasis looks at differences between the expected and observed ORs, when comparing the combined ORs derived from two variants with the ORs for the two variants individually [129]. Given that two variants contribute to the risk to develop a disease. If a case-control study identifies variant A to have an OR = 2 and variant B to have an OR = 3, it is expected that for individuals carrying both variants the effects of these two variants are additive

resulting in an OR = 5 and thus a 5 fold increased risk for disease. If the observed OR = 10 for instance, statistical epistasis is present as the additive character of the two variants can not be observed. An example for statistical epistasis derives from a study of Keck *et al.* [130]. In their study, multiplicative effects of exonic polymorphisms in the *CRHR1* and *AVPR1B* genes in a case-control sample for PD were identified. While the p-values for association with PD were 0.001 for rs878886 in *CRHR1* and 0.015 for rs28632197 in *AVPR1B*, the p-value for the multilocus effect was 0.00057. Several statistical methods for the detection of epistasis exist. Unfortunately, statistical models do not necessarily correlate with biological models for epistasis and thus, performing a statistical test and concluding from statistical interactions that also biological interactions exist, is not possible [131].

1.1.5.4 Gene-environment interactions

The aim of GWASs is to discover genetic variants that have direct main effects on disease susceptibility [71]. However, variants that show no association with the investigated disease may nevertheless contribute to the risk to develop the disease through hidden gene-environment interactions. If the risk for a disease is increased in individuals, who both carry a susceptible genotype to the condition under investigation and are exposed to a specific environmental risk factor compared to individuals, who are either exposed to the risky environment or are genetically susceptible, gene-environment interactions are present (Figure 5).

A variety of environmental risk factors for mental disorders exist, including maternal stress during pregnancy, maternal substance abuse during pregnancy, low birth weight, birth complications, deprivation of normal parental care during infancy, childhood physical maltreatment, childhood neglect, premature parental loss, exposure to family conflict and violence, stressful life events involving loss or threat, substance abuse, toxic exposures and head injury [132]. The hypothesis of genetic moderation assumes that differences between individuals regarding their genomic background result in differences in their vulnerability or resilience to environmental factors so that environmental influences are considered to be only contributing to a disease and not causing it [133]. Pharmacogenetic studies confirm the hypothesis of genetic moderation as different genotypes were identified to be associated with different drug responses [134].

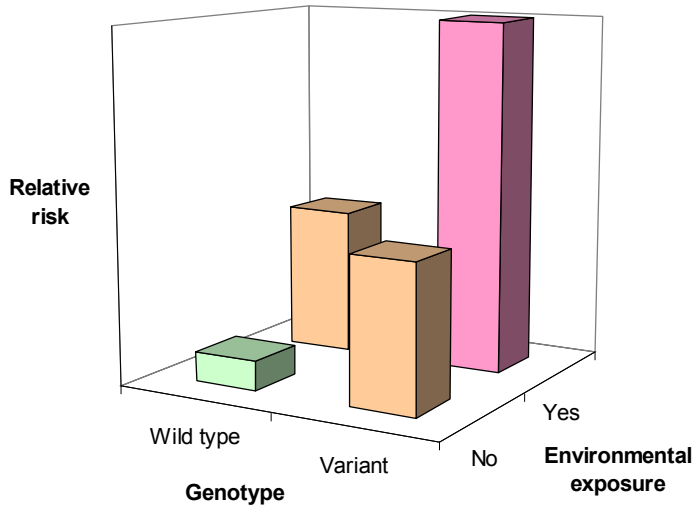


Figure 5 Model of gene-environment interaction.
Both the genotype is dichotomous (carrier versus non-carrier of a genetic variant) and the exposure with a specific environmental factor is dichotomous (exposed versus non-exposed). Out of the four possible genotype-exposure combinations only individuals, who are both carrying the risk variant and have been exposed to a distinct environment (pink), show an increased relative risk to develop the disease. Adapted from Hunter [135].

The first study describing the role of the genotype in moderating the effects of environmental factors to develop a mental disease was published in 2002. In this study, the hypothesis that a functional SNP in the promoter region of the *MAOA* gene, which encodes for a neurotransmitter-metabolizing enzyme, influences the effects of maltreatments in childhood on violent behaviour was investigated [136]. The results showed that children who were maltreated in childhood and carried the genotype conferring to low levels of *MAOA* expression suffered more often from conduct disorder, antisocial personality and adult violent crime than maltreated children with the genotype associated with higher *MAOE* expression.

The contribution of genetic and shared and individual environmental factors to the phenotype can be estimated by comparing the disease concordance rate between MZ and DZ twins. As already mentioned in section 1.1.2, a higher concordance rate in DZ twins indicates that shared environmental factors contribute to a high degree to the investigated disease while genetic factors play a major role in disease susceptibility when the concordance rate is increased in MZ twins. Family-based designs such as sibling pairs or case-parent designs allow also the estimation of the proportion to which the phenotype is determined by environmental factors.

Studies, which investigate the penetrance of a genetic variant, can also unravel the extent of environmental contribution to a phenotype. Changes in penetrance over time lead to the suggestion that changes in the environment have been occurred. One example for a changed penetrance comes from a study investigating 333 women, who carry the *BRCA1* gene [137]. The penetrance was increased in more recent birth cohorts indicating that the influence of novel environmental and lifestyle factors increase the risk for breast cancer.

A challenge for the investigation of gene-environment interactions is the need for large sample sizes. It has been estimated that for the detection of joint effects, a four times larger sample is needed compared to the sample size that would be necessary to detect the main effects of genetic or environmental factors individually [138]. Thus, studies that were designed to capture the main effects of individual factors are mostly underpowered to examine interactions.

1.2 Genetics of anxiety and mood disorders

1.2.1 Classification and clinical characteristics of anxiety disorders

Anxiety disorders (ADs) are among the most common psychiatric disorders in children [139], adolescents [140] and adults with a combined life-time prevalence of 28.8% [141]. Substantial impairment [142], loss in work productivity [143] and the use of primary care services [144] indicate a high burden of this disorder.

The International Statistical Classification of Diseases and Related Health Problems (ICD-10) Manual of the World Health Organization (WHO) classifies ADs in phobic anxiety disorders and other anxiety disorders. In the first class, anxiety is only caused by well-defined situations or objects. Phobic anxiety disorders are further subdivided into three forms: the first is agoraphobia, the fear of having a panic attack in places or in situations where the person feels unable to escape. Social phobia is associated with low self-esteem and fear of criticism leading to the avoidance of social situations such as talking in the public, visiting parties or meeting friends. Specific or isolated phobia is the most prevalent form of ADs, with a lifetime prevalence of 12.5% in the general population [141]. Such phobias are restricted to highly specific situations, including particular animals, closed spaces, height, flying, darkness or the seight of blood.

Other ADs, which are further subdivided into PD and generalized anxiety disorder (GAD), are characterized by a sudden appearance of anxiety without any particular environmental stimuli. The main symptoms of GAD are worries and anxieties in the daily life. Fears that the patient itself or relatives become ill or have an accident are often

expressed. The quality of life of patients suffering from permanent anxieties is severely impaired, often leading to suicidal ideation [145]. PD is characterized by recurrent and unexpected attacks of intense fear that last for several minutes. Additionally, four of the following somatic symptoms occur: heart palpitation, accelerated heart rate, xerostomia, tremor, dyspnea, nausea, sweating. PD is often accompanied by agoraphobic behaviour. ADs are characterized by early onset, the tendency to have a chronic course and high comorbidity with each other [146] and other psychiatric disorders such as bipolar disorder [147] and depression [148]. The median age at onset was reported to be six years in an adolescent population-based study [140] and eleven years in a population-based sample of adults [141]. Furthermore, age at onset varies within the different forms of ADs. Specific phobias often occur in childhood [141], social phobia in early and late adolescence [149], PD with or without agoraphobia in late adolescence and early adulthood [150] and GAD in early adulthood [141]. Children and adolescents, suffering from ADs, have an increased risk to suffer either from the same or a different form of AD in adulthood [151].

1.2.2 Genetic epidemiology of ADs

There is substantial evidence that ADs are familial and heritable. Family studies showed that the risk to develop PD was significantly increased (7.9 - 17.3%) in first-degree relatives of PD patients compared to relatives of unaffected subjects [152-155]. Furthermore, the risk of first-degree relatives of PD patients to develop PD is depended on the age at onset of the PD patient. An onset before the age of 20 increases the disease risk by 17 fold, whereas the risk is only increased by 6 fold when PD occurred after the age of 20 [156]. Familial aggregation has also been observed for GAD and phobias [157].

Twin studies revealed a 2 – 3 times higher concordance rate for PD in MZ twins compared to DZ twins [158]. The Virginia Adult Twin study, comprising 5000 twin pairs, reported a genetic proportion of variance in liability to different ADs as followed: PD 0.28, GAD 0.23, animal and situational phobia 0.24 and social phobia 0.1 [44]. The remaining variance in liability could be attributed to individual-specific environmental influences. These heritability estimates indicate that ADs are moderately heritable compared to other psychiatric disorders such as schizophrenia and bipolar disorder, with heritability scores of 0.67 and 0.62 respectively [46].

1.2.3 Genetic studies in ADs

Many linkage studies have been undertaken to identify chromosomal regions that harbour susceptibility genes for mainly PD, yielding a variety of potential risk loci on chromosomes 1q [159], 2q [160], 4q31-q34 [161], 7p [162], 9q [163], 12q [164], 13q [165], 14q [166] and 22q [167]. However, linkage analyses have shown little consistency.

The candidate gene literature in ADs is extensive, but only a few associations could be replicated [168,169]. The three most studied genes are catechol-O-methyltransferase (*COMT*), brain-derived neurothropic factor (*BDNF*) and solute carrier family 6 member 4 (*SLC6A4*, also known as serotonin transporter *5-HTT*) [170]. The *COMT* gene encodes an enzyme which methylates catecholamines, including the neurotransmitters dopamine, epinephrine and norepinephrin, leading to its degradation. A SNP in this gene causes a substitution of valine (Val) with methionine (Met) at codon 158 (Val158Met polymorphism) which is associated with a 40% [171] or even three to four fold lower COMT activity [172]. A meta-analysis of six case-control studies identified the Val allele and the associated increased activity of the COMT enzyme as possible risk factor for PD, although the effects are different in European and Asian populations and female-specific [173].

The *BDNF* gene is known to be involved in the actions of the serotonergic [174], glutamatergic [175] and dopaminergic [176] neurotransmitter systems. High gene expression levels in the hippocampus and the involvement in neurotransmission suggested that this gene might contribute to different neuropsychiatric traits such as anxiety [177]. The most investigated variant in *BDNF* leads to an amino acid exchange from Val to Met at codon 66 (Val66Met polymorphism). In a mouse model, a higher anxiety-like behaviour could be observed in homozygous Met/Met BDNF mice when exposed to stress (results not normalized by the antidepressant fluoxetine) [178]. In contrast, a meta-analysis of seven case-control studies on AD in humans could not find an association between the BDNF Val66Met polymorphism and AD [179].

To date, one of the most frequently studied polymorphisms in psychiatric research has been the serotonin transporter length polymorphic region (5-HTTLPR) within the *SLC6A4* loci. This length polymorphism in the promoter region results in two alleles which are distinguishable by a 44 bp insertion (long allele) or deletion (short allele) [180]. The short allele is associated with a decreased *SLC6A4* gene expression and reduced serotonin uptake [180,181]. While an association between the short allele and anxiety traits could be observed in healthy subjects [182], a meta-analysis, comprising ten case-control

studies on PD, could not identify an association between the 5-HTTLPR polymorphism and this class of AD [183].

So far, the number of GWASs for anxiety-related phenotypes is very small, with only two publications reporting data from a GWAS for PD. The first study identified two genome-wide significant SNPs in the transmembrane protein 16B (*TMEM16B*) and plakophilin 1 (*PKP1*) gene which, however, could not be replicated in another Asian PD sample [184,185]. In a second GWAS, the transmembrane protein 132D (*TMEM132D*) locus was identified as a possible novel candidate gene for PD [4]. A two SNP haplotype was associated with PD in three independent German samples. These risk alleles were also associated with higher *TMEM132D* messenger RNA (mRNA) expression in human postmortem cortex samples. Moreover, in a mouse model of extremes in trait anxiety [186], a higher *Tmem132d* expression in the anterior cingulated cortex was observed in high versus low anxiety animals. This collected evidence suggests that an increase in *TMEM132D* expression may be associated with the development of pathological anxiety. The initial results could be replicated in additional European cohorts, assembled as part of the Panic International Consortium (PanIC) [187].

1.2.4 Clinical characteristics of major depressive disorder

In 2030, major depressive disorder (MDD) is projected to be one of the three leading causes of burden of disease [188]. MDD, which is also known as unipolar depression, is a complex and severe psychiatric disorder associated with high rates of morbidity and mortality. The lifetime prevalence in the general population is estimated to be 13 – 16%, with a ~ 1.8 higher rate in women than in men [148,189]. The risk of mortality due to suicide is substantial and the lifetime risk estimates are 3.4%, with a higher risk for males than females (7% versus 1%) [190].

MDD is characterized by a variety of symptoms along several biological and psychological systems, including the endocrine, affective, immune, autonomic-vegetative and cognitive system [191]. The ICD-10 classification system defines a major depressive episode by depressed mood and/or loss of interest in usual activities, combined with three or more additional symptoms, including appetite and sleep disturbances, loss of energy, feelings of worthlessness and guilt, lack of emotional reactions, diminished ability to think or concentrate, suicidal thoughts and ideation for a duration of two weeks or longer. Depending on the number of depressive episodes, the ICD-10 manual further subdivides MDD into single depressive episode and recurrent depressive disorder.

An early age at onset is one characteristic of MDD. The National Comorbidity Survey (NCS), which studies the prevalence of mental disorders in the US population, reported that the half of all lifetime cases of mood disorders starts at age 14 and 75% by 24 years [141]. Furthermore, it has been shown that age at onset is associated with course and severity of disease. Earlier age at onset is associated with greater illness burden compared to patients developing the disease later which is manifested in a greater number of depressive episodes with increased severity of symptoms, social and occupational impairment, poorer quality of life, greater medical and psychiatric comorbidity, a more negative outlook, increased suicidal ideation and attempted suicide [192].

MDD is highly comorbid with other mental disorders. The NCS study showed that 75% of depressed individuals suffer at least from one other mental disorder [193]. These patients are reported to have significantly poorer psychosocial functioning and poorer recovery rates compared to patients with only depression diagnosis [194]. The strongest comorbidity is with AD which is present in 50% of individuals suffering from MDD [195].

1.2.5 Genetic epidemiology of MDD

The number of epidemiological studies investigating MDD is limited as, especially in older studies, both unipolar and bipolar depressed individuals were included into the investigations. The latter subjects are characterized by changes between manic and depressive phases. Only five family studies, five twin studies and none adoption study meet the stringent inclusion criteria for a meta-analysis, although two out of three adoption studies pointed out that genetic factors influence the risk for MDD [45].

Family studies indicated that MDD is aggregated in families as first-degree relatives of depressed individuals have a 2.84 increased risk to develop MDD itself compared to relatives of a healthy individual [45]. In addition, some other family studies suggest that prepubertal onset depression is largely influenced by environmental factors, while depression, which occurs after adolescence, is more influenced by genetic components [196,197].

Twin studies argued for a substantial genetic and unique environmental component of MDD, while shared environmental factors contribute to little or no extent to the phenotype. The meta-analysis of five twin studies estimated the heritability to 37% and the influence of the individual specific environment to 63% [45].

1.2.6 Genetic studies in MDD

In the past, numerous linkage studies were performed to identify susceptibility regions for MDD. Unfortunately, results obtained in these mainly smaller linkage studies with less than 100 affected individuals were inconsistent [198-200]. In contrast, studies investigating several hundreds of affected individuals identified two genomic regions that present evidence for linkage with MDD. In the Genetics of Recurrent Early-Onset Major Depression (GenRED) study, which comprises 415 affected sibling pairs, genome-wide linkage was observed on chromosome 15q [201]. Two independent large-scale studies, the European-US Depression Network study [202] and a study from Utah [203], provided also support for chromosome 15q as susceptibility region for MDD. The second region, identified in the combined first and second wave European-US Depression Network study [204] and another independent MDD sample [205], was on chromosome 3p.

Candidate gene studies suggested *SLC6A4, HTR2A, BDNF, TPH2, APOE, GNB3* and *MTHFR* to be associated with MDD [206]. Among possible MDD candidate genes, serotonin-related genes are the most investigated ones as changes in serotoninergic function regarding receptor density, metabolism and reactivity have been observed in MDD, and pharmacological effects of serotoninergic drugs in MDD are well documented [207]. Furthermore, genes controlling other biochemical substances such as gamma-aminobutyric acid (GABA), glutamate, and norepinephrin have been investigated as these substances are also assumed to play a major role in MDD [208]. Overall, only a small number of candidate genes could be replicated in independent studies. Therefore, several meta-analyses were conducted to increase the power to detect risk alleles for depression. One of the largest meta-analysis in depression included 183 studies covering 393 genetic variants in 102 possible candidate genes [209]. This comprehensive study identified variants in the genes *APOE* (apolipoprotein E), *GNB3* (guanine nucleotide-binding protein ß-3), *MTHFR* (methylene tetrahydrofolate reductase), and *SLC6A4* to influence the susceptibility for depression. For the latter gene, it was found that carriers of the short allele of the 5-HTTLPR polymorphism have a slightly increased risk for MDD with an OR of 1.11 [209], while another meta-analysis reported that neither the genotype of the serotonin transporter alone nor in interaction with stressful life events is associated with an increased risk for MDD [210].

In the last years, several GWASs for MDD have been published. None of these studies reported genome-wide significant results and their findings were difficult to replicate [211-216]. In 2011, a GWAS identified *SLC6A15* as novel susceptibility gene for MDD [5]. rs1545843 risk allele carrier status was associated with a decreased *SLC6A15* gene

expression in the hippocampus. This brain region, which modulates the hypothalamic-pituitary-adrenocortical (HPA) axis, is dysregulated in depressed patients [217]. Furthermore, the same SNP showed an association with reduced hippocampal volume and reduced hippocampal neuronal integrity. Environmental factors such as chronic stress were also associated with a reduced *Slc6a15* gene expression in stress-susceptible compared to stress-resilient mice (1.9 fold in CA1 region).

1.3 Detection of genetic variants using NGS

1.3.1 Introduction into NGS technologies

The introduction of DNA sequencing by chain termination coupled with size separation of the obtained DNA fragments via gel electrophoresis was one of the most important technological developments in the field of genetics. This sequencing technique, which was invented by the Nobel laureate Sanger and which is thus also referred to as Sanger sequencing method, uses chemically modified dideoxynucleotide triphosphates (ddNTPs) to terminate the DNA strand during chain elongation [218]. Since the development of automated capillary sequencers, which are based on the Sanger sequencing method, identification of whole genome sequences became feasible [219].

Although this first-generation sequencing method led to a revolution in biological science, time and costs remained the major limitations for large-scale sequencing studies for single research laboratories. For instance, in the framework of the Human Genome Project estimated 300 million dollars and ten years of time were necessary to generate the first draft of the human reference genome [29-31]. NGS methods, which are often also referred to as massively parallel sequencing (MPS) techniques, lead to a drastic decrease in costs per sequenced base pair and increase in speed compared to capillary sequencing [220,221]. Using these NGS technologies, an individual genome can be sequenced within a few weeks with costs currently less than 10,000 dollars [222]. Although NGS technologies were originally designed to sequence whole genomes (whole genome sequencing), they can also be used in a more restricted way investigating only the entity of exonic regions in the genome, also called exome sequencing [223,224], or clearly defined genetic regions, for instance single genes. Targeted re-sequencing approaches are the most feasible ones for single laboratories [225].

1.3.2 NGS workflow

The optimal NGS design is dependent on the individual scientific question which is aimed to be answered by the sequencing project. If a study is focussed on the detection of rare variants in one ore more single genes, whole genome sequencing is not indicated, but targeted re-sequencing is sufficient. When performing a targeted re-sequencing experiment the question which genomic region to sequence arises. Genes, which were identified in earlier performed GWASs to be associated with the investigated phenotype, are often chosen for sequencing studies. While in GWASs only a limited number of common variants within a gene can be detected directly or indirectly via LD, NGS offers the possibility to identify, in theory, all common and rare genetic variants in the investigated gene and thus, NGS is often used as follow-up experiment of GWASs. Furthermore, results from the literature can also influence the choice of candidate genes. In case of unknown candidate genes for a disease, sequencing the entire exome offers an unbiased approach.

After defining the target region, enrichment of DNA in this region has to be performed using either methods based on polymerase chain reaction (PCR) or hybridization-based technologies which are carried out either in solution or on microarrays. The best-suited enrichment technology depends on the size of the target region and the sample size. In general, enrichment using PCR is attractive for studies with small target sizes and small sample sizes, while hybridization-based methods are indicated for whole exome sequencing projects or targeted re-sequencing studies with large sample sizes [222]. For whole genome sequencing studies the target enrichment step can be omitted.

In the next step, genomic or enriched target DNA has to be prepared for the sequencing experiment. The preparation process is dependent on the used sequencing platform. Currently, three major NGS platforms are available namely Illumina with the Genome Analyzer, Life Technologies with the sequencing by oligonucleotide ligation and detection (SOLiD) sequencer and Roche with the 454 FLX sequencer. In this thesis, both targeted re-sequencing projects were performed using the SOLiD sequencer. Therefore, the description of the further sequencing workflow in the following sections is based on this sequencing platform. An overview of other NGS technologies is given elsewhere [226].

1.3.2.1 Library preparation

Library preparation is the first of two sections of sample preparation for sequencing and starts with randomly breaking the enriched target DNA into smaller fragments comprising hundreds of base pairs using ultrasound treatment. The sheared DNA has 5' and/or 3' protruding ends which are converted into blunt-ends with repair enzymes. After the end-repair step, the DNA is purified using column or bead purification. Then, P1 and P2 adaptors, which are short double-stranded oligonucleotides, are ligated to the ends of the end-repaired and purified DNA. Adaptor-ligated DNA is run on a size selection gel and the ligation products with the desired size are extracted (150 – 200 bp). The size-selection step is recommended as DNA fragments with similar lengths result in more uniform library amplification due to the fact that shorter fragments are more likely to be amplified than longer ones which introduce an amplification bias. In addition, size selection removes possible adaptor dimers. The size-selected fragments are quantified and, if necessary, amplified with library PCR Primer 1 and 2. These primers are specific to the P1 and P2 adaptor sequences so that only DNA fragments to which a P1 or P2 adaptor was ligated are amplified. After amplification, the PCR products are column or bead-purified and the final library is quantified.

1.3.2.2 Bead enrichment

In the second section of sample preparation DNA obtained after library preparation is clonally amplified onto DNA beads using emulsion PCR (ePCR) [227]. Clonal amplification, which results in a population of identical templates, is required as most NGS technologies have not been designed to detect single fluorescence signals emitted from one single DNA template in the sequencing reaction. The emulsion is created by mixing an oil phase with an aqueous phase so that droplets of aqueous phase are distributed in the oil phase. The aqueous phase contains all components which are required to perform the ePCR, including library template DNA, P1 and P2 primers, nucleotides, DNA polymerase and P1 DNA beads. It is desired that each droplet in which the clonal amplification takes place contains one single bead and one single template DNA so that monoclonal micro-reactors are present. During ePCR, which results in approximately 30,000 copies of template per bead, the P1 adaptor of the template DNA binds complementary to the P1 adaptor which is attached to the DNA beads. Therefore, only P1 adaptor carrying DNA templates are amplified. After ePCR, the emulsion is broken by adding alcohol and the beads are washed to remove the oil. Then, a bead enrichment step is necessary as not all beads carry a template. For this reason,

enrichment beads with a P2 adaptor are added to the amplified and purified beads. These enrichment beads can only build a complex with beads carrying a template as the template harbours the P2 adaptor sequence. Enrichment results in a two phase solution with enrichment beads with or without template beads in the upper layer and beads without templates at the bottom. The upper layer is extracted and denatured to isolate the template beads from the enrichment beads. In the last step, the template beads are extended with a Bead Linker which is necessary for deposition on the sequencing glass slide.

1.3.2.3 SOLiD sequencing by ligation

The actual sequencing reaction is initiated by the binding of a sequencing primer to the P1 adaptor of the templated beads. Fluorescently labelled two-base-encoded probes hybridize to the complementary sequence of the template and a DNA ligase joins the probe to the primer [228]. Two-base-encoded probes are oligonucleotide sequences in which the two first bases are associated with a particular fluorescence dye, while the following six bases are either degenerated (N = 3) or universal (N = 3). The incorporation of a probe leads to the emission of a fluorescence signal which can be imaged by the sequencer to determine the ligated probe [229]. After fluorescence imaging, the three universal nucleotides and the fluorescent dye are cleaved and the next probe can be incorporated into the growing strand. The SOLiD cycle, which comprises ligation, fluorescence imaging and cleavage, is repeated nine more times. As from each incorporated probe only the first two bases are associated with a colour, the remaining three degenerated nucleotides are not detectable. For this reason, the primer extension product is removed from the template strand after ten ligation cycles, and a second primer binds to the P1 adaptor of the template however, in comparison to the position where the first primer bound, shifted by one base position. Five rounds of primer re-setting, each with ten ligation cycles is required to determine a DNA fragment. Through the primer reset process, every base is interrogated in two independent ligation reactions by two different primers.

The generated sequencing reads are denoted in color space which is the ordered sequence of the detected fluorescence signals from the five ligation rounds (5 x 10 cycles). Thus, a direct determination of the DNA sequence is not possible. To convert the sequence of the colour calls into the DNA sequence, information about the two-base code is necessary (Figure 6). Assuming that two bases encode for a green fluorescence signal and that the last base of the sequencing primer is an A, then a green color can

only be observed when the first incorporated base is a C. If the next fluorescence signal is blue, the following base has to be a C as only the combination CC decodes for blue and so on. Raw reads in color-space are subjected to quality control procedures and the reads, which survived this quality control, are aligned to a known color-space reference sequence to identify genetic differences between template and reference.

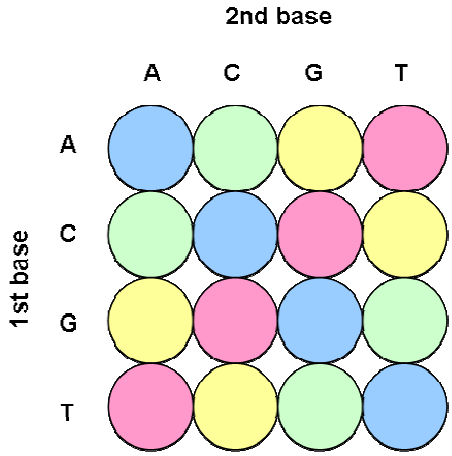


Figure 6 Two-base encoding scheme.
In the two-base encoding scheme, four dinucleotide sequences are associated with a fluorescence color respectively. Adapted from Metzker [226].

1.4 Aims of the investigation

Complex traits are known to be influenced by a complex interplay between multiple genes and environmental factors. To identify the genes which contribute to the manifestation of a disease, linkage and association studies have been extensively carried out. The understanding of the genetic architecture of a disease could be beneficial for its diagnosis, treatment and prognosis.

In the last years, GWASs have identified thousands of common genetic variants associated with common complex diseases (see the NHGRI Catalog of Published Genome-Wide Association Studies). Unfortunately, the heritability explained by these factors is small [3]. The most prominent explanations for the case of missing heritability are the contribution of multiple rare variants, structural variants, gene-gene interactions and/or gene-environment interactions to a phenotype. With regard to the role of rare variants, several studies have been shown that a combination of common and rare genetic variants contributes to the risk to develop common diseases [230-232]. Therefore, the debate whether the CDCV hypothesis postulating that common variants with small effects are disease causing or the CDRV theory, which suggests that multiple rare variants with larger effects contribute to a disease, is true should not longer seen as an either/or debate, but should be shifted to the question to which extent common and rare variants are involved in the aetiology of a disease.

1.4.1 Genetic and *in silico* functional characterization of the *TMEM132D* locus

In 2011, a GWAS identified *TMEM132D* as novel candidate gene for PD [4]. *TMEM132D* on chromosome 12 encodes a single-pass type 1 membrane protein belonging to the TMEM132 protein family. A haplotype, containing two common variants within the intronic regions of the *TMEM132D* locus, was associated with PD in three independent samples. This initial finding could be replicated in additional European cohorts derived from the PanIC consortium [187]. In addition, three independent common SNPs in *TMEM132D* were associated with the severity of anxiety symptoms in patients with a number of different primary psychiatric disorders. Susceptibility genes for a disease, which were identified in GWASs, are amenable for follow-up studies as genes harbouring common variants are likely to contain also rare variants [6]. Therefore, the question arises whether besides common variants also rare variants play a role in the susceptibility to anxiety-related phenotypes.

In order to answer this question, a pooled targeted re-sequencing experiment using the SOLiD sequencing platform was performed. Until the development of NGS technologies,

the investigation of rare variants was complicated. The combination of low frequency and low LD between common and rare variants makes this type of genetic variation unsuitable for the analysis with microarrays which are used in GWAS [233]. The exonic regions and the exon-intron boundaries of the *TMEM132D* gene were re-sequenced in 300 AD patients and 300 controls which were arranged in four DNA pools. A subset of the detected common and rare variants was re-genotyped in all 600 subjects using matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry on the Sequenom platform. Validated variants were then subjected to association analysis. For common variants with a MAF > 5%, PLINK was used for association testing. To test whether rare variants are associated with AD, these were collapsed and the combination of variants was subjected to association analysis.

In addition to the detection of associations between common and/or rare variants and AD, it was also focussed on the assessment of functional relevance of these variants. Functional characterization is important as genetic association does not necessarily have to mean causation. Unfortunately, the function of the TMEM132D protein is still unknown so that only *in silico* annotation could be performed in order to detect possible functional relevant variants.

1.4.2 Genetic and experimental functional characterization of the *SLC6A15* gene

In 2011, another GWAS suggested SLC6A15 as candidate gene for MDD [5]. A common variant about 600 kb downstream of this gene was identified to decrease *SLC6A15* gene expression in lymphoblastoid cell lines as well as hippocampus. Furthermore, the same SNP showed an association with reduced hippocampal volume in patients with depression and reduced hippocampal neuronal integrity in healthy controls. In this thesis, it was focussed on the question whether common and rare variants within the *SLC6A15* locus lead to the symptoms of MDD.

To accomplish this aim, the whole *SLC6A15* locus, including 10 kb up- and downstream of the gene, was re-sequenced in 400 MDD patients and 400 controls which were combined in eight DNA pools. After the pooled targeted re-sequencing run, which was also performed on the SOLiD sequencer, a subset of the detected variants was individually re-genotyped using MALDI-TOF mass spectrometry. Validated variants were then tested for association with MDD using the same statistical methods as for the TMEM132D study (section 1.4.1).

As it is known that *SLC6A15* encodes an amino acid transporter which is highly expressed in the brain [7], amino acid uptake is a measurable property that associates

with function. Therefore, functional characterization was not restricted to computational annotation, but experimental assessment of functional relevance of genetic variants could also be performed. Non-synonymous variants were incorporated into the SLC6A15 protein using site-directed mutagenesis. Then, the uptake of proline was measured in a SLC6A15 uptake assay which was developed by the Chemical Genomics group of Felix Hausch at the Max Planck Institute of Psychiatry.

2. MATERIALS AND METHODS

2.1 Recruitment and sample characterization

2.1.1 AD sample

300 patients (122 males, 178 females) from the Anxiety Disorders Outpatient Clinic of the Max Planck Institute of Psychiatry (MPIP) in Munich were included in the study. The mean age was 37.9 ± 11.9 years (males: 37.3 ± 10.9 years, females: 38.3 ± 12.5 years). Of the included patients, 84.7% presented PD with or without agoraphobia as their primary psychiatric diagnoses (Table 3). The diagnosis was ascertained by trained psychiatrists according to the Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV criteria. All patients underwent the Structured Clinical Interviews for DSM-IV (SKID I and II). AD due to medical or neurological condition or a comorbid Axis II disorder was an exclusion criterion. All patients underwent a thorough medical examination including EEG, ECG and detailed hormone laboratory assessment. The mean age at onset was 27.3 ± 11.4 years (males: 26.9 ± 11.1 years, females: 27.6 ± 11.6 years). Severity of anxiety and depression was measured using the 14-item Hamilton Anxiety Scale (HAM-A), the 21-item Hamilton Depression Scale (HAM-D) and the Bandelow Panic and Agoraphobia Scale (PAS) [234-236]. A mean HAM-A score of 23.9 (SD: 9.4) and a mean HAM-D score of 13.8 (SD: 6.2) indicated moderate anxious and low depressed patients. The severity of panic and agoraphobia symptoms was moderate to high with a mean PAS score of 29.8 (SD: 9.4). Ethnicity was recorded using a questionnaire for nationality, mother language and ethnicity of the subject itself and all four grandparents. All included patients were Caucasian, 82.7% of German origin. 17.3% were from countries other than Germany, mostly from countries in Eastern Europe and Mediterranean countries.

300 controls were recruited at the MPIP. Individuals were selected randomly from a Munich-based community sample and screened for the absence of the following psychiatric disorders: mood disorders, psychotic symptoms, AD, alcohol dependence, drug abuse, obsessive/compulsive disorders, Post-Traumatic Stress Disorder (PTSD), dissociative disorders, somatoform disorders and eating disorders using the Munich version of the Composite International Diagnostic Interview (M-CIDI) [237]. M-CIDI is an updated version of the World Health Organization's CIDI version 1.2 (WHO-CIDI) which incorporates questions to cover DSM-IV (American Psychiatric Association) and ICD-10 (WHO) diagnostic criteria. Only subjects with a negative life-time history of the above-

mentioned disorders were included in the study, representing a group of individuals from the general population who has never been mentally ill. The controls were matched for ethnicity (using the same questionnaire as for patients), age and gender. All controls were Caucasian with 91.3% of German origin and the remaining controls mostly from Eastern or Western European countries.

Table 3 Demographic and clinical characteristics of the investigated AD sample. Adapted from Quast *et al.* [238].

Characteristics	Cases	Controls
N	300	300
Sex		
male	40.7% (122)	44.7% (134)
female	59.3% (178)	55.3% (166)
Age (SD)	37.9 (11.9)	38.8 (11.0)
Diagnosis	PD with agoraphobia 66.0% PD without agoraphobia 18.7% Social phobia 8.7% GAD 4.0% Agoraphobia 1.7% Specific phobia 0.7%	none
HAM-A (SD)	23.9 (9.4)	NA
HAM-D (SD)	13.8 (6.2)	NA
PAS (SD)	29.8 (9.4)	NA

SD, standard deviation; PD, panic disorder; GAD, generalized anxiety disorder; HAM-A, Hamilton Anxiety Scale Score; HAM-D, Hamilton Depression Scale Score; PAS, Bandelow Panic and Agoraphobia Scale Score; NA, not assessed

2.1.2 MDD discovery sample

400 unipolar depressed patients (166 males, 234 females) from the Munich Antidepressant Response Signature (MARS) of the MPIP were subjected to the study [239,240]. The mean age was 46.9 ± 12.9 years (males: 46.6 ± 11.7 years, females: 37.2 ± 13.7 years). Of the included patients, 88.0% suffered from recurrent depressive disorder, 12.0% presented a single depressive episode (Table 4). Patients were included in the study within 1 – 5 days of admission to the clinic and diagnosed from trained psychiatrists according to the DSM-IV criteria. Exclusion criteria were the presence of alcohol or substance abuse or dependence (including eating disorders with laxative abuse), comorbid somatization disorder, and depressive disorders due to medical or

neurological conditions. The mean age at onset was 31.9 ± 12.2 years (males: 33.7 ± 11.9 years, females: 30.7 ± 12.2 years). Severity of depression and anxiety was measured using the HAM-D and HAM-A score. Patients fulfilling the criteria of a HAM-D score ≥ 18 for recurrent depressive disorder or ≥ 20 for single depressive episode and an age at onset ≤ 55 were included in the study. Ethnicity was recorded using a self-report questionnaire for nationality, mother language and ethnicity of the subject itself and all four grandparents. All included patients were Caucasian, 78.0% of German origin. 22.0% were of European descent, mostly from Eastern and Western Europe.

400 controls were selected randomly from the Munich general population and screened for the absence of the in section 2.1.1 mentioned psychiatric disorders. The included individuals were matched for ethnicity, age and gender. All controls were Caucasian, 91.8% of German origin.

2.1.3 MDD replication sample

905 patients (294 males, 611 females) were recruited at the MPIP and psychiatric hospitals in Augsburg and Ingolstadt, both in Germany. The mean age was 50.9 ± 13.8 years (males: 50.0 ± 13.6 years, females: 51.6 ± 13.8 years). All patients suffered from recurrent major depression (Table 4) [213,241]. The mean age at onset was 36.0 ± 13.9 years (males: 36.7 ± 14.0 years, females: 35.6 ± 13.9 years). Diagnoses were ascertained by trained psychiatrists according to the DSM-IV criteria using the WHO Schedule for Clinical Assessment in Neuropsychiatry (SCAN, version 2.1). Only patients over 18 years with at least two moderately severe depressive episodes were included in the study. Individuals with a positive life-time history of the following psychiatric disorders were excluded: presence of manic episodes, psychotic symptoms, presence of intravenous drug abuse and depressive symptoms only secondary to alcohol or substance abuse or to medical illness or medication. Ethnicity was recorded using a self-report sheet for nationality, mother language and ethnicity of the subject itself and all four grandparents. All included patients were Caucasian, 89.5% of German origin. 10.5% were from countries other than Germany, mostly from countries in Eastern and Western Europe.

1029 controls, matched for age, gender and ethnicity to the patient sample, were selected randomly from a Munich-based community sample and screened for the absence of anxiety and affective disorders using the Composite International Diagnostic-

Screener. Only individuals without the above mentioned disorders were included, representing a healthy control group regarding anxiety and depression.

All studies were approved by the local ethics committee of the Ludwig-Maximilians-University (LMU) in Munich. Written informed consent was obtained from all individuals.

Table 4 Demographic and clinical characteristics of the MDD discovery and replication sample. Adapted from Quast *et al.* [242].

Characteristics	Discovery sample		Replication sample	
	Patients	Controls	Patients	Controls
N	400	400	905	1029
Sex				
male	41.5% (166)	41.8% (167)	32.5% (294)	32.7% (336)
female	58.5% (234)	58.3% (233)	67.5% (611)	67.3% (693)
Age (SD)	46.9 (12.9)	46.9 (15.1)	51.1 (13.8)	50.7 (13.9
Diagnosis				
recurrent depressive disorder	88.0% (352)	none	100.0% (905)	none
single depressive episode	12.0% (48)		-	
HAM-D (SD)	27.4 (5.0)	NA	NA	NA
HAM-A (SD)	25.5 (8.2)	NA	NA	NA
age at onset (SD)	31.9 (12.2)	NA	36.0 (13.9)	NA

SD, standard deviation; HAM-D, Hamilton Depression Scale Score; HAM-A, Hamilton Anxiety Scale Score; NA, not assessed

2.2 DNA enrollment

On enrollment in the study, up to 40 ml blood were drawn from each patient or control. DNA was extracted from whole blood using a standardized extraction procedure (Puregene whole blood DNA-extraction kit, Gentra Systems Inc). Genomic DNA was quantified using picogreen based fluorometry and adjusted to 50 ng/µl.

2.3 DNA amplification and pooling design

For the TMEM132D study, equal amounts of genomic DNA from 50 patients or 50 controls were combined in a DNA pool (Figure 7). Thus, six pools comprising DNA from AD patients and six pools including DNA from control subjects were prepared for the following amplification step. In order to amplify all nine exons of the *TMEM132D* gene on chromosome 12 (129,556,271-130,388,212, hg19) and the boundaries from exon to

intron, eight oligonucleotide primer pairs were designed which covered target regions of approximately 5 kb in length (40 kb in total). Primer sequences are listed in the supplement section (Table S1). DNA amplification of each amplicon in each pool was performed using Long Range-PCR (LR-PCR) resulting in 96 individual PCR reactions (twelve pools x eight amplicons). LR-PCR was carried out in a 96 well plate in a reaction volume of 50 µl. Each PCR reaction contained 200 ng genomic DNA, 0.8 µM of each primer, 300 µM of each deoxynucleotide and 2.5 units of LongAmp Taq DNA polymerase (New England Biolabs (NEB)). The cycling protocol was as follows: 94°C for 3 minutes, then 94°C for 30 seconds, 61°C for 40 seconds and 65°C for 5 minutes for 30 cycles. The final extension was carried out at 65°C for 10 minutes.

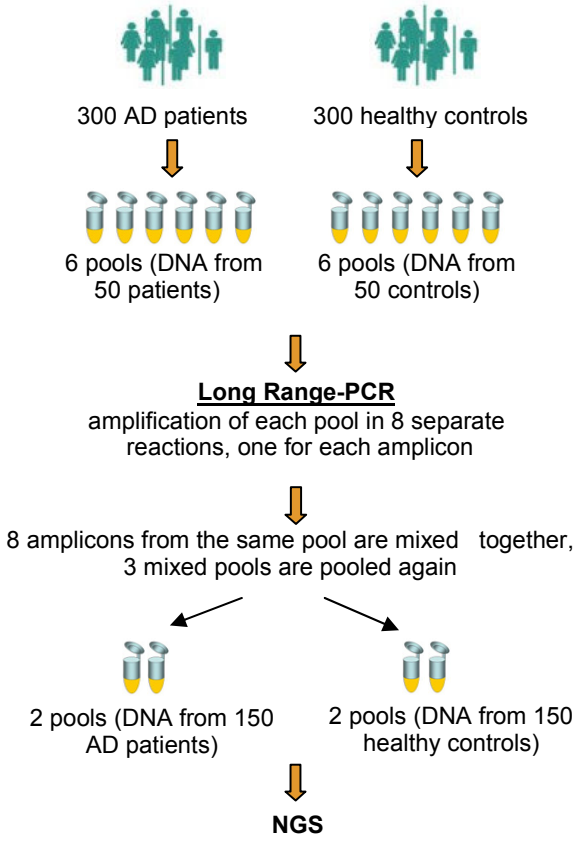


Figure 7 Pooling design used for the TMEM132D project.
Combining 50 AD patients and 50 controls in a DNA pool reduces the number of individual PCR reactions from 4800 (600 individuals x 8 amplicons) to only 96. After amplification mixing the DNA fragments which derived from the same pool and combining three mixed pools reduces the number of pools for the following NGS experiment to only four.

In order to check the success of the amplification step, all PCR products were loaded on a 0.8% agarose gel. In the next step, the concentrations of the obtained DNA fragments

were measured using the Qubit High Sensitivity dsDNA Kit on the Qubit 1.0 fluorometer (Invitrogen). All eight amplicons which were generated from the same pool were mixed together at an equimolar ratio resulting in twelve DNA pools. To remove undesired DNA fragments and primer dimers, the pools were loaded on a 0.8% agarose gel and the corresponding DNA band was extracted using the QIAQuick Gel Extraction kit (Qiagen). The DNA concentration of the purified DNA pools was measured using the same quantification method as described above. In the last step, equimolar amounts of three pools were mixed together (Figure 7). Finally, two pools containing amplified DNA from 150 AD patients and two pools comprising DNA fragments from 150 controls were subjected to preparation for NGS.

For the SLC6A15 project, the same DNA amplification method and pooling strategy was used with exception of the following changes. Genomic DNA from 50 MDD patients or 50 controls was arranged in pools at an equimolar ratio resulting in eight case pools and eight control pools for amplification. Eleven primer pairs generating amplicons between 2 and 11 kb were designed to amplify the whole *SLC6A15* locus on chromosome 12 (85,253,267-85,306,608, hg19) and 10 kb upstream and downstream of the gene (70 kb in total). With exception of a 3.5 kb intronic region for which a working primer pair could not be designed, the whole gene was covered. The sequences of the used primer pairs are listed in the supplement section (Table S2). Each LR-PCR reaction was carried out in a volume of 25 µl and contained 100 ng DNA, 0.8 µM of each primer, 300 µM of each deoxynucleotide and 2.5 units of LongAmp Taq DNA polymerase. Depending on the melting temperature of the primer pair and the amplicon length the cycling protocol was as follows: 94°C for 3 minutes as initial denaturation, then 94°C for 30 seconds, 59 - 61°C for 40 seconds and 65°C for 2.5 - 11 minutes for 32 cycles. The final extension was carried out at 65°C for 10 minutes. After mixing all eleven DNA fragments from the same pool, equal amounts of DNA from two pools (two case pools and two control pools respectively) were combined. Performing this pooling strategy four pools comprising DNA from 100 MDD patients and four pools including DNA from 100 controls were generated for the subsequent NGS experiment.

2.4 Library preparation and bead production

For the TMEM132D study, fragment library preparation was performed according to the SOLiD 3+ System Library Preparation Guide (P/N 4442697 Rev. A; 01/2009). DNA amounts between 860 and 1000 ng from each of the four generated pools were used as

starting material. After adaptor ligation and size selection, seven cycles of up-scale PCR resulted in library concentrations between 364 and 983 pg/µl. Quality control as well as quantification of the libraries was performed using the high sensitivity DNA Kit on the Agilent 2100 Bioanalyzer.
According to the SOLiD 3+ System Templated Bead Preparation Guide (P/N Part Number 4442695 Rev. A 10/2009), bead production and enrichment was carried out manually as a full-scale reaction. From each library, a final concentration of 0.5 pM was subjected to the ePCRs. Finally, the quality and the amount of the templated beads were checked with a workflow analysis run (WFA). 87 to 110 million beads per pool were deposited on a 4-well sequencing slide which was physically separated into four equal compartments, one compartment for each pool. The sequencing run was performed using SOLiD Opti sequencing chemistry for a single F3 Tag with a read length of 50bp. Primary data analysis was done on the SOLiD 3+ instrument with default settings. The generated .csfasta and .qual files were exported for further analysis.

For the SLC6A15 project, two sequencing runs were performed as in the first run the coverage of two amplicons was low. For the first run, DNA amounts between 300 and 500 ng from each of the eight pools were subjected to barcoded standard fragment library preparation with size selection, following the SOLiD 4 System Library Preparation Guide (P/N 4443045 Rev. B. 04/2010). After 10-12 cycles of up-scale PCR, amplified libraries were quantified with a TaqMan assay.
Equal amounts of all eight barcoded libraries were combined in a multiplex and subjected to bead production and enrichment which was carried out using the EZ bead system according to the SOLiD 4 System Templated Bead Preparation Guide (P/N Part Number 4442695 Rev. A 10/2009). For the ePCR, a final concentration of 1.5 pM of the generated multiplex was used as DNA template. Enriched template beads were quality checked and quantified performing a workflow analysis run (WFA). Each, approximately 476 million beads were deposited on two full slides. The SOLiD sequencing run was performed using SOLiD TOP Fragment Barcoding Sequencing chemistry for a single F3 Tag with a read length of 50 bp. Primary data analysis was done on the instrument with default settings. For further analysis .csfasta and .qual files were exported.
For the second SLC6A15 run, library and bead preparation were performed in the same way as for the first run. For library preparation, 500 ng DNA from each pool were used as input material. Library amplification was carried out with 7 cycles of upscale-PCR. Amplified libraries were quantified using the Qubit High Sensitivity dsDNA Kit on the

Qubit 1.0 fluorometer. For emulsion PCR, a final concentration of 0.8 pM of the multiplexed library consisting of equal amounts of eight DNA pools was used as DNA template. Beads were deposited on two lanes of a six lane Flow Chip for sequencing on the SOLiD 5500xl Sequencer using SOLiD FWD SR sequencing chemistry for a single F3 Tag with a read length of 75bp.

2.5 Variant validation

In order to confirm variants which were detected in the NGS experiment, individual re-genotyping using MALDI-TOF mass spectrometry on the Sequenom platform (San Diego, USA) was performed. The MassARRAY Assay Designer software was used for primer selection, multiplexing, assay design and mass-extension for generating primer extension products. For genotype calling, the MassARRAY Typer 3.4 software was used. All Sequenom experiments were performed at the Helmholtz Zentrum in Munich, Germany.

From the detected variants in the *TMEM132D* locus, a subset of 151 variants was re-genotyped in the 300 AD patients and the 300 control subjects. Variants were selected based on the following criteria: synonymous and non-synonymous variants, variants with high and low ORs (OR > 2 and OR < 0.5) and variants located in 5' and 3' untranslated regions (UTRs), transcription factor binding sites (TFBSs) and evolutionary conserved regions.

69 variants in the *SLC6A15* gene, which were detected in the two NGS runs, were individually re-genotyped in the MDD discovery sample comprising 400 patients and 400 controls. Selection criteria for variants to re-genotype were the same as for the TMEM132D study. Non-synonymous variants, which could be confirmed in the discovery sample, were also re-genotyped in the MDD replication sample, including 905 depressed patients and 1092 controls. In addition, 22 non-synonymous variants from the Exome Sequencing Project (ESP) database (Exome Variant Server, http://evs.gs.washington.edu/EVS/) were re-genotyped in the replication sample. Currently, the ESP database incorporates exome sequencing data from 6,503 individuals from European, African and Asian populations (accessed in April 2013).

2.6. Functional characterization

2.6.1 *In silico* functional analysis

An easy and convenient possibility to assess the functional relevance of a genetic variant is *in silico* annotation. Several computational tools were used in this work. In order to

identify variants in regulatory regions of the gene, all detected variants were mapped to Encyclopedia of DNA Elements (ENCODE) TFBSs (http://genome.ucsc.edu) [243] which were determined by chromatin immunoprecipitation sequencing (Chip-Seq). In addition, variants were mapped to ENCODE/Duke DNaseI hypersensitivity sites in several brain regions, including cerebellum, frontal cerebrum and frontal cortex (http://genome.ucsc.edu) [243]. Furthermore, it was investigated whether variants were located in micro RNA (miRNA) regulatory target sites in the 3' UTRs of genes, predicted by TargetScanHuman 5.1 (http://targetscan.org) [244]. For the identification of variants with putative effects on splicing such as disrupting existing exonic splicing enhancer (ESE) or silencer (ESS) motifs or creating new splice sites, FASTSNP was used (http://fastsnp.ibms.sinica.edu.tw/) [245]. Variants located in genomic regions with high degree of nucleotide conservation were identified using PhastCons [246] and PhyloP [247]. The computational annotation tools Sorting Intolerant From Tolerant (SIFT) [248], PolyPhen2 [249] and Panther [250] were used to predict the effects of non-synonymous coding variants on the function of the gene product, based on amino acid conservation across different organisms.

2.6.2 Experimental functional analysis

2.6.2.1 Site-directed mutagenesis

While the function of the TMEM132D protein is still unknown, SLC6A15 was identified to transport neutral amino acids into predominantly neuronal cells [7]. Thus, nine validated non-synonymous coding variants within the long isoform of the *SLC6A15* locus could be experimentally tested for alterations on protein function. The first step of the experimental functional characterization was the insertion of the human SLC6A15 cDNA (clone name IRAKp961L15168Q; ImaGenes) into the pEGFP-C1 vector (Clonetech) using restriction enzymes specific to BglII and SalI sites. In the generated eGFP-hSLC6A15 construct, hSLC6A15 cDNA is fusioned with the sequence encoding the enhanced green fluorescent protein (eGFP) at the N-terminus. Variants of the eGFP-hSLC6A15 construct, containing one of the nine non-synonymous variants, were generated using the site-directed mutagenesis technique (Table 5).

The incorporation of the genetic variant into the hSLC6A15 cDNA was done via PCR with oligonucleotides containing the desired mutation. All primer sequences are listed in the supplement section (Table S3). PCR was carried out in a reaction volume of 50 µl with 10 ng of plasmide DNA, 125 ng of each primer, 200 µM of each deoxynucleotide and 1.0 unit of Phusion High Fidelity DNA Polymerase (NEB). The cycling protocol was

as follows: 98°C for 30 seconds, then 98°C for 30 seconds, 55°C for 1 minute and 72°C for 9 minutes for 25 cycles. The final extension was carried out at 72°C for 15 minutes. After amplification, 1 µl restriction endonuclease DpnI (NEB) was added and the samples were incubated for 1 h at 37°C. 5 µl of each PCR product were transformed into E.*Coli* DH5α cells performing a heat shock for 1 minute at 42°C. Transformed cells were suspended in LB Medium, plated on agar plates containing the antibiotic kanamycin (50 µg/ml) and incubated over night at 37°C. The next day, three colonies were picked from each agar plate and inoculated in LB Medium with kanamycin (50 µg/ml) overnight at 37°C. Plasmid DNA was isolated and purified using the HiYield Plasmid Mini Kit (Real Biotech Corporation). Success of the site-directed mutagenesis was verified by Sanger sequencing of the plasmid DNA.

Table 5 Overview of the mutant plasmids created by site-directed mutagenesis. Adapted from Quast *et al.* [242].

Mutant name	Nucleotide exchange	Amino acid exchange	Position in protein
hSLC6A15 T49A	A → G	Thr → Ala	49
hSLC6A15 K227N	G → C	Lys → Asn	227
hSLC6A15 A400V	C → T	Ala → Val	400
hSLC6A15 L421P	T → C	Leu → Pro	421
hSLC6A15 I500T	T → C	Ile → Thr	500
hSLC6A15 N591D	A → G	Asn → Asp	591
hSLC6A15 A601T	G → A	Ala → Thr	601
hSLC6A15 E684D	G → C	Glu → Asp	684
hSLC6A15 G710R	G → A	Gly → Arg	710

2.6.2.2 ^{3}H proline uptake assay

Human embryonic kidney (HEK) 293 cells were transfected with wild type or one of the nine mutated plasmids using Lipofectamine (Invitrogen). HEK293 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Gibco) containing 10% of fetal calf serum (FCS) and 5% Penicillin/Streptomycin at 37ºC in a humidified incubator (5% CO_2). For transfection, 24 µg of each plasmid were mixed with 1.5 ml complete DMEM. In a second preparation, 60 µl Lipofectamine and 1.5 ml DMEM were mixed together. Lipofectamine contains lipid subunits which can build liposomes entrapping genetic material. These liposomes easily merge with membranes and thus inject their content into cells. After five minutes of incubation at room temperature, the lipofectamine/DMEM mixture was added to the plasmid preparation and incubated for further 20 minutes at

room temperature. Finally, the plasmid/lipofectamine mixture was added to the cells. The following day, transfected cells were detached, counted and plated in 96 well plates. In order to increase the adhesion of the cells, wells were pre-coated with Poly D-Lysine (PDL). After two hours, PDL was removed and wells were washed with Phosphate Buffered saline (PBS) solution. Each well contained 60,000 transfected cells in 150 µl complete DMEM respectively.

In preparation of the cellular uptake experiment, DMEM was removed from each well with five washing steps using the Deep-Well Microplate Washer ELx405 from Biotek. Then, to the 100 µl remained from the washing steps, 50 µl of the non-labeled neutral amino acid L-proline were added to each well, followed by adding 50 µl proline with a final concentration of 20 nM which was labeled with radioactive tritium (^{3}H) (Perkin Elmer). The cellular uptake of ^{3}H proline was measured in dependence of different concentrations of the non-labeled L-proline (final concentrations of 3 µM, 12 µM, 48 µM, 195µM, 781 µM, 3.1mM, 12.5 mM and 50.0 mM) so that for wild type and each mutant eight uptake measurements were performed respectively. After 10 minutes of incubation, the proline/^{3}H proline mixture was removed with six washing steps. To lyse the cells, 10 µl of NaOH (1M) were added to each well and the plate was shaked for five minutes. Finally, 200 µl scintillation cocktail (Perkin Elmer), which amplifies the radioactive signal, was added to each well. The plate was shaked for 20 minutes and then incubated for two hours in the dark. ^{3}H proline uptake was measured using the Wallac MicroBeta luminescence counter (Perkin Elmer). Transfected HEK cells containing wild type or altered eGFP-hSLC6A15 plasmids were measured in triplicates.

2.6.2.3 Fluorescence imaging

While the cellular uptake assay offers the possibility to investigate whether point mutations in the *SLC6A15* gene affect the function of the amino acid transporter, information about possible quantitative changes are missing. In order to assess altered protein levels and/or a changed cellular localization of the protein, fluorescence imaging was performed. For that purpose, HEK293 cells were plated on cover slips which were precoated with PDL. After one day, the medium was removed and cells were fixed with 4% paraformaldehyde (PFA). Excess of PFA was removed performing two washing steps with PBS solution. The cover slips were mounted on slides using 4µl mounting medium containing DAPI which stains cell nuclei blue. After few hours of incubation, samples were analysed at the confocal microscope.

2.7 Statistical analysis

2.7.1 Statistical analysis of NGS data

In the framework of statistical analysis of the generated NGS data, which was performed in collaboration with the Biostatisticians of the research group, raw reads were subjected to the following quality control (QC) procedure: reads with more than four colors who have a phred-like quality value of <= 10 were trimmed before the fifth color of insufficient quality occurs. Trimmed reads, which were shorter than 30 colors, were excluded from further analysis. For further details of the QC process, see Altmann et al. [251]. QC filtered reads were aligned to the reference sequence on chromosome 12 of the human genome (NCBI Build 36.1 for TMEM132D and the first SLC6A15 run and GRCh37 for the second SLC6A15 run) using the Burrows-Wheeler aligner (BWA) version 0.5.7 [252] and the Short Read Mapping Package (SHRiMP) aligner version 2.2.0 [253]. Four mismatches between sequencing read and reference were allowed.

VipR, which was developed to call variants in pooled samples, was used for variant detection in both projects [251]. Briefly, vipR counts how often the minor allele (MA) of a variant at a given base position occurs in a DNA pool and compares the number with the MA calls at the same position obtained from the remaining pools. Then, vipR assesses the likelihood that the base calls originate from sequencing errors. Only if the likelihood is sufficiently small, then the altered allele at that position is reported as a true genetic variant. This variant caller is freely available at http://sourceforge.net/projects/htsvipr/. A prerequisite for the inclusion of a sequenced base into variant calling was a minimum in coverage at that position of at least 5,000 in each pool.

After variant calling, annotation of these variants was performed using the Annotation variant (ANNOVAR) tool, freely available at http://www.openbioinformatics.org/annovar/ [254].

2.7.2 MAF correlation

Correlation between MAFs obtained in the NGS experiment and in the re-genotyping stage was assessed using SPSS version 18.0. The same tool was used to compare the MAFs derived from the validation experiments and the MAFs denoted in the public available ESP database for European Americans.

2.7.3 Association testing

2.7.3.1 Association analysis of common variants

In both studies of this thesis, association of common variants with AD or MDD was tested using the PLINK tool (http://pngu.mgh.harvard.edu/~purcell/plink/) [255]. Variants who were not validated in the re-genotyping experiment or who had a MAF below 5%, a genotyping rate below 90% and a deviation from Hardy-Weinberg equilibrium at a significance level below 0.05 were excluded from this analysis. Applying these selection criteria, 32 out of 36 common variants within the *TMEM132D* gene and two out of 2 re-genotyped SNPs within the *SLC6A15* locus were included into allelic association testing. Furthermore, individuals with a genotyping rate < 90% were excluded as well. An alpha level of 0.05 after correction for multiple testing using the Benjamini-Hochberg method correcting for all tested common variants was considered statistically significant.

2.7.3.2 Association analysis of rare and/or putatively functional variants

Power calculations were performed using Quanto version 1.2.3 (http://hydra.usc.edu/gxe/) [256]. Single-marker testing tools such as PLINK are unsuitable for association analysis involving rare variants as the power to detect an association with a single rare variant is low [97,257]. An alternative approach is to test whether a combination of multiple rare and/or putatively functional relevant variants is associated with disease. In more detail, the presence of minor alleles (PMA) and the sum of minor alleles (SMA) was compared between patients and controls. For the investigation of the PMA, an individual who harboured any of the variants included into the collapsed marker set, independent of the total number of variants and the number of altered alleles per variant (heterozygous or homozygous carriers of a variant) was encoded with 1. Subjects without any of the tested variants were encoded with 0. To test for differences in the SMA between cases and controls, homozygous carriers of a variant were encoded with 2 and individuals with one altered minor allele were encoded with 1. Subjects in which the variant was not present were encoded with 0. The SMA is thus a quantitative or numeric variable, while the PRA has a qualitative or categorical character as it gives only information whether an individual carriers one or more variants or not.

In the TMEM132D study, three different SNV sets were tested for differences in the PMA and the SMA between AD patients (N = 300) and controls (N = 300) or between PD patients (N = 252) and controls. The first set contained all SNVs with a MAF <= 1% (N = 66). In the second set, all coding SNVs (N = 25) were included. Finally, all non-synonymous variants and all coding variants with predicted effects on splicing (N = 20)

were tested for association. For the putatively functional SNVs, no MAF cut-off was used. Statistical significance was assessed using linear regression models implemented in R and following 1000 permutation of the case-control status.

In the SLC6A15 project, the tested SNV set contained all non-synonymous variants which were discovered in the NGS experiment (N = 9). This marker set was investigated for differences in the PMA and SMA between MDD patients (N = 1305) and controls (N = 1429) of the combined sample. Statistical significance was assessed using independent samples t-test for the SRA and contingency tables for the PRA in SPSS version 18.0.

For both studies, p-values were not corrected for the multiple comparisons. The level of significance was set to 0.05 for these tests.

2.7.3.3 Population stratification

To correct for allele frequency differences between cases and controls due to different ethnic backgrounds, the method of genomic controls was performed using PLINK. Genomic control is based on the idea that unlinked genetic markers which are distributed across the human genome can be used to test for the proportion of allelic diversity between cases and controls [258]. If population stratification is present then the Chi square association statistic is increased by a factor which is proportional to the extent of population stratification. This factor is referred to as lambda (λ) and is defined as the quotient between the median observed Chi square association statistics across all tested SNPs between cases and controls and the theoretical median under the null-hypothesis of no stratification [259,260].

For genomic control, a set of unlinked variants, which was genotyped in 272 cases and 300 controls of the AD sample using the Illumina Human-1 100k, Illumina HumanHap300-Duo and Illumina Human610-Quad BeadChips, was tested for differences in allele frequency. SNPs with a genotyping rate < 98% and/or showing deviation from Hardy-Weinberg equilibrium at an error level of below 10^{-5} were excluded from the analysis. Rare variants with a MAF < 5% were also excluded, resulting in 287,515 SNPs for the analysis. Included SNPs were genotyped in all individuals. For all SNPs passing the quality criteria as mentioned above, Chi square values for allelic case-control association were calculated. The distribution of these Chi square values was compared with the theoretically expected Chi square distribution.

In addition, principal components analysis (PCA) was performed to assess population differences [261,262]. This method attempts to structure individuals in subgroups according to their similarity of allele frequencies of tested unlinked markers and to

compare each subgroup. If population stratification is not present only one subgroup exists which comprises both cases and controls. Possible outlying individuals were visualized in a multidimensional scaling (MDS) plot.

2.7.4 Statistical analysis of ^{3}H proline uptake assay

The maximal uptake of ^{3}H proline, which occurs in the absence of antagonists (non-labeled L-proline), and the IC_{50}, which is the concentration of non-labelled proline where the inhibition of the ^{3}H proline uptake is 50%, were assessed using Sigma Plot. Differences in the mean ^{3}H proline uptake between wild type and mutant were assessed using general linear models in SPSS version 18.0. Transfection efficiency was included as covariate into the analysis. An alpha level of 0.05 after correction for multiple testing using the Bonferroni method was considered statistically significant.

3. RESULTS

3.1 Results of the genetic and *in silico* functional characterization of the *TMEM132D* locus

3.1.1 Data from the pooled targeted re-sequencing experiment

Re-sequencing all nine exons and exon-intron boundaries of the *TMEM132D* gene (40 kb in total) in four pools, each comprising DNA from 150 AD patients or healthy controls, resulted in 328 million raw reads with a length of 50 bp (Table 6). Thus, approximately 82 million reads per DNA pool were generated. From the obtained raw reads 200 millions (61.0%) had a sufficient quality and survived the QC procedure. Of these, 93.2% could be mapped to the reference sequence (NCBI Build 36.1, UCSC hg18). For each DNA pool, an average of 46.6 million reads was thus subjected to the subsequent variant calling step. Overall, 56.9% of the generated raw reads were mappable and could be used for further analyses.

Table 6 Number of reads generated in the TMEM132D re-sequencing experiment.
Reads were mapped to the reference sequence using the BWA aligner. The numbers of the raw, quality filtered and mappable reads are given in millions. Adapted from Quast *et al.* [238].

DNA pool	raw reads	QC filtered reads	mappable reads
Cases 1	82.2	48.9	45.1
Cases 2	77.5	48.6	45.4
Controls 1	86.5	53.3	50.1
Controls 2	81.7	49.2	45.8
total	327.9	200.0	186.4

QC, quality control

The average sequencing depth in this NGS run was approximately 50,000 fold per base position per DNA pool and therefore theoretically 330 fold per subject (150 patients or controls per pool) and 165 fold per allele (300 alleles per pool). 90.0% of the whole sequenced region, 96.0% of the exonic region and 100.0% of the protein coding sequence were sequenced with a mean coverage > 5,000 in each pool. Hence, 10.0% of the sequenced DNA bases were excluded from subsequent variant calling procedure. While the average coverage per pool was 50,000 fold, the actual coverage at a given base position showed large variations, ranging from 300 to 400,000 fold. Especially the ends of the amplicons were more often covered by sequencing reads than sequences in

the middle of the amplified DNA fragment. Although the coverage distribution within a pool was uneven, the pattern of sequencing depth was highly conserved across all four DNA pools (Figure 8).

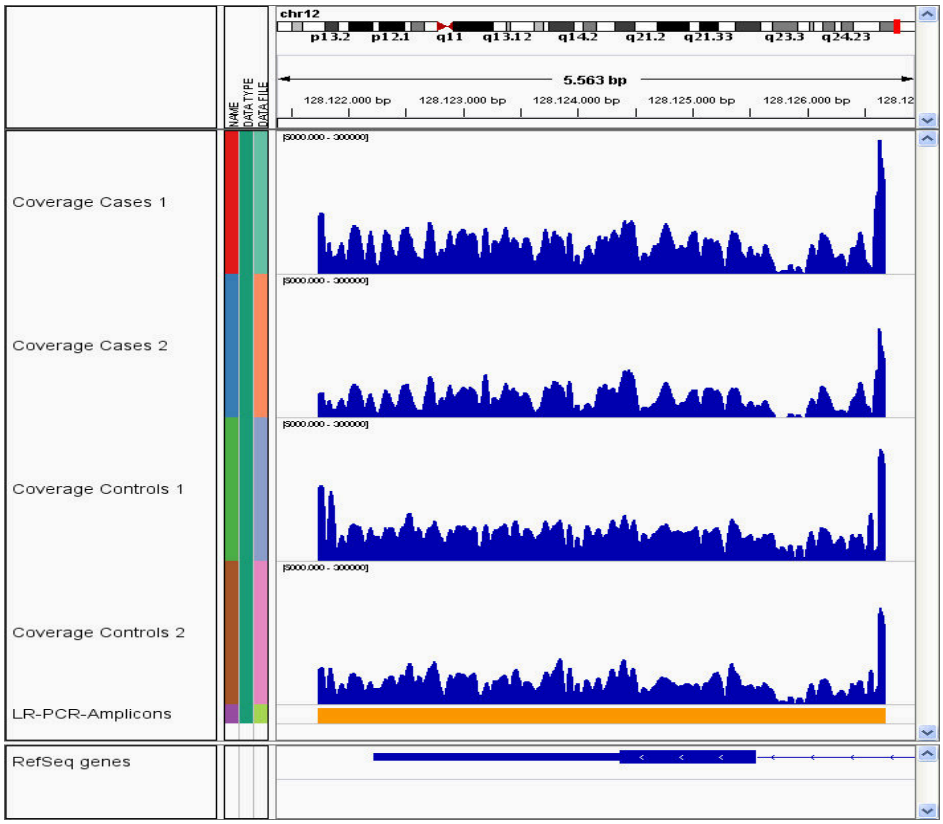


Figure 8 Coverage distribution within and across the four sequenced DNA pools.
The orange box denotes the position of the 5 kb amplicon which covers exon nine of the *TMEM132D* gene. The individual coverage of each pool is plotted from 5,000 to 300,000 reads per base. Regions with a coverage less than 5,000 reads/base in any pool were excluded from variant calling. Due to the similar coverage distribution in all eight sequenced DNA fragments of a pool, only one single amplicon is shown. The figure was generated using the Integrative Genomics Viewer (IGV) [263]. Taken from Quast *et al.* [238].

3.1.2 Detection of variants in *TMEM123D*

Using the VipR algorithm [264], 371 genetic variants in the *TMEM132D* locus on chromosome 12 were identified. Of these, 53.9% (N = 200) were novel and had not been previously reported in the dbSNP137 database (accessed in April 2013). For three of the variants, which were at positions of the genome at which a genetic variant was already published in dbSNP137, discrepancies regarding the minor allele were observed. For instance, rs12815188 was listed in dbSNP137 as a base exchange from G to C, whereas the NGS as well as the subsequent re-genotyping experiment showed a G to A

conversion. Thus, the number of novel variants in *TMEM132D* increased from 200 to 203 SNVs. Of the 371 detected SNVs, 151 (40.7%) have also been identified in the 1000 Genome Project database (April 2012 release) [35].

120 and thus approximately one third of the detected variants in *TMEM132D* were extremely rare, with MAFs <= 0.5%. Of these variants, only 22.5% (N = 27) were recorded in dbSNP137, whereas almost all common SNPs (95.2%) with MAFs above 5% were previously detected.

From the variants, which were identified in the NGS experiment, 8.1% (N = 30) were in the coding region of the gene, 4.0% in the 3' UTR and only one variant (0.3%) was located in the 5' UTR. Hence, the vast majority (87.6%, N = 325) of the variants in *TMEM132D* were located outside the exonic regions. Of the 30 coding SNVs, 17 were non-synonymous variants which lead to amino acid exchanges in the gene product (Figure 9). From these missense variants, only five have been previously published in dbSNP137.

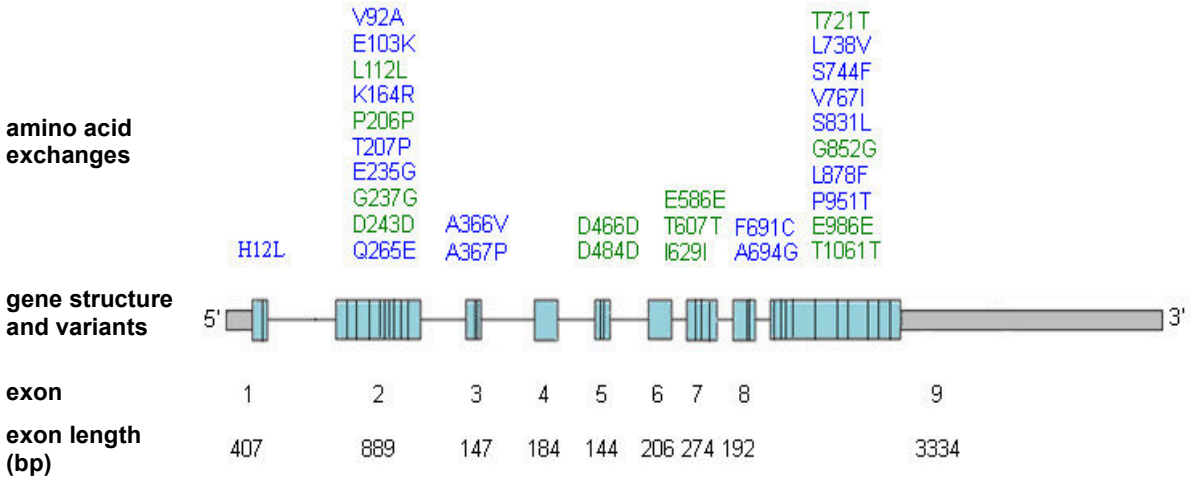


Figure 9 Detected variants in the coding region of the *TMEM132D* locus.

The blue boxes denote the coding regions of the gene, the grey boxes the untranslated regions. A genetic variant is indicated as vertical black line. Variants annotated in green are synonymous SNVs, variants annotated in blue lead to amino acid exchanges in the protein. Note that exons but not introns are drawn to scale. Adapted from Quast *et al.* [238].

3.1.3 Variant validation using MALDI-TOF mass spectrometry

From the subset of genetic variants (N = 150), which was subjected to individual re-genotyping using MALDI-TOF mass spectrometry as an independent method, 144 variants (95.4%) could be sucessfully re-genotyped (Figure 10). Out of these variants, 119 (83.2%) could be confirmed as polymorphic, including 25 coding variants (Table 7).

For validated variants with a MAF < 15.0% (N = 101), the correlation between the MAFs estimated in the NGS experiment and verified by individual re-genotyping was excellent (*r = 0.974*). However, the comparison of the MAFs for validated variants with a MAF higher than 15.0% (N = 18) showed a low correlation (*r = 0.435*).

From the 25 validated coding variants in table 7, seven were only present in patients and three were only observed in controls. Interestingly, six of the variants restricted to patients and one SNV exclusively found in controls were neither previously reported in dbSNP137 nor present in any sequences from the 1000 Genomes Project or the ESP database, currently incorporating 6503 individuals (accessed in April 2013). Fur further information, see table 7 in which these SNVs were highlighted in bold.

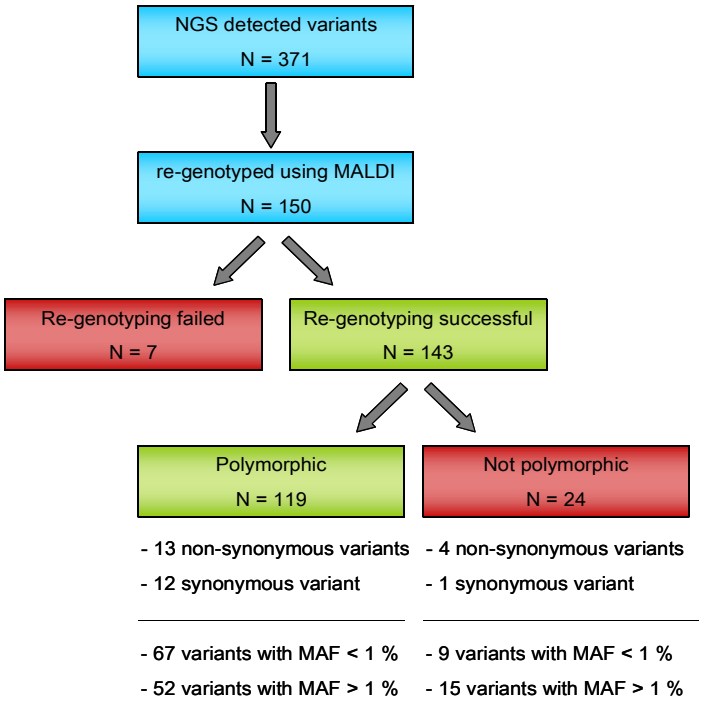


Figure 10 Validation of variants in *TMEM132D* using MALDI-TOF mass spectrometry.
Denoted MAF was estimated from NGS discovery stage. Adapted from Quast *et al.* [238].

3.1.4 *In silico* functional annotation of coding and non-coding variants in *TMEM132D*

In silico functional analysis of all validated coding variants identified 14 variants to have putatively effects on splicing (Table 7). Nine variants were predicted to create new ESE or ESS motifs, four to disrupt already existing splice sites and one variant to disrupt an ESS motif and to create a new ESE site simultaneously. For nine of the non-synonymous SNVs, the evolutionary nucleotide conservation tools phastCons and phyloP showed consistent results and predicted four variants to be located in

evolutionary conserved regions (Table 7). The prediction of the functional relevance of the validated non-synonymous variants was inconsistent across the used evolutionary amino acid conservation tools SIFT, PolyPhen2 and Panther. For none of the tested variants (N = 13), a deleterious effect on protein function was predicted in all three tools (Table 7). Subjecting all 371 detected variants to TFBS and miRNA analysis identified 18 variants in predicted TFBS (Table 8). Two of these SNVs, were located in exon 3. Variants disrupting miRNA target sites were not observed.

Table 7 *In silico* functional characterization of validated coding variants in *TMEM132D*. Adapted from Quast *et al.* [238].

				MAF	MAF					NT conservation		AA conservation		
SNV	Location on chr12	SNV in dbSNP137	Alleles	Case (%)	Con (%)	OR[1]	Location within gene	AA exchange	Splicing analysis[2]	Phast Cons[3]	PhyloP[4]	SIFT[5]	Panther[6]	Poly Phen2[7]
chr12_128953803	130387850	rs142888394	T > A	1,17	0,88	1,34	exon 1	His / Leu	new ESE site	-	-	-	-	+
chr12_128751001	130185048	rs151244831	A > G	1,85	1,88	0,98	exon 2	Val / Ala	ESE site broken	-	+	+	-	-
chr12_128750969	130185016	-	C > T	0,17	np	**Case**	exon 2	Glu / Lys	no effect	+	-	-	-	++
chr12_128750942	130184989	rs143257185	G > A	0,33	0,17	2,00	exon 2	Leu / Leu	no effect	+				
chr12_128750565	130184612	rs140064887	C > T	1,17	1,84	0,63	exon 2	Gly / Gly	no effect	-				
chr12_128750547	130184594	rs76591377	G > A	0,17	0,17	0,99	exon 2	Asp / Asp	new ESE site	-				
chr12_128750483	130184530	-	G > C	0,17	np	**Case**	exon 2	Gln / Glu	new ESS site	+	+	-	-	-
chr12_128581575	130015622	-	G > A	0,17	np	**Case**	exon 3	Ala / Val	no effect	-	+	-	-	-
chr12_128581573	130015620	rs144926631	C > G	np	0,17	Con	exon 3	Ala / Pro	no effect	-	-	-	-	-
chr12_128260081	129694128	rs12815188	G > A	np	0,17	Con	exon 5	Asp / Asp	no effect	+				
chr12_128260027	129694074	rs145352969	G > A	0,17	0,34	0,49	exon 5	Asp / Asp	new ESE site	-				
chr12_128132422	129566469	rs77363876	C > T	1,33	1,17	1,14	exon 7	Glu / Glu	new ESE site	-				
chr12_128132359	129566406	rs79031518	C > T	12,00	10,37	1,16	exon 7	Thr / Thr	no effect	-				
chr12_128132293	129566340	rs60962336	G > A	1,34	1,17	1,15	exon 7	Ile /Ile	new ESE site	+				
chr12_128129066	129563113	-	G > C	0,67	np	**Case**	exon 8	Ala / Gly	no effect	-	-	-	+	-
chr12_128125510	129559557	rs10773594	T > C	28,02	33,99	0,82	exon 9	Thr / Thr	ESS site broken	-				
chr12_128125461	129559508	-	A > C	0,17	np	**Case**	exon 9	Leu / Val	ESE site broken	+	-	-	-	++
chr12_128125442	129559489	-	G > A	np	0,17	**Con**	exon 9	Ser / Phe	new ESE site/ESS site broken	+	+	-	+	++
chr12_128125374	129559421	rs73159540	C > T	8,67	11,32	0,77	exon 9	Val / Ile	no effect	+	+	-	-	-
chr12_128125181	129559228	rs140762080	G > A	0,84	1,37	0,61	exon 9	Ser / Leu	no effect	-	-	-	-	-
chr12_128125117	129559164	rs148302846	T > C	0,33	np	Case	exon 9	Gly / Gly	no effect	-				
chr12_128125039	129559086	rs555131	C > A	1,89	3,23	0,58	exon 9	Leu / Phe	new ESE site	-	-	-	-	-
chr12_128124822	129558869	-	G > T	0,17	np	**Case**	exon 9	Pro / Thr	new ESS site	+	+	+	-	++
chr12_128124715	129558762	rs77778748	C > T	1,34	2,37	0,56	exon 9	Glu / Glu	new ESS site	+				
chr12_128124490	129558537	rs74344050	G > A	0,84	1,34	0,62	exon 9	Thr / Thr	ESS site broken	-				

SNV, single nucleotide variant; chr, chromosome; MAF, minor allele frequency; Con, control; np, not polymorphic; OR, odds ratio; AA, amino acid; ESE, exonic splicing enhancer; ESS, exonic splicing silencer; NT, nucleotide

Location on chr12 is according to the February 2009 Human Reference Sequence (UCSC Genome Browser). Known SNVs are recorded in the dbSNP137 database.

[1] Case: SNV only in cases; Con: SNV only in controls; bold: SNV neither in dbSNP137 nor in the re-sequencing databases of the 1000 Genomes and Exome Sequencing Project [2] FastSNP [3] ANNOVAR, phastCons 44-way alignment [4] ANNOVAR, phyloP alignment, restricted to non-synonymous variants: + conserved (score > 0.95), - non-conserved (score < 0.95) [5] SIFT (sorting intolerant from tolerant): - tolerant (score > 0.05); + possibly damaging (score < 0.05) [6] Panther: - unlikely functional effect (pdeleterious < 0.5); + possibly damaging (pdeleterious > 0.5) [7] PolyPhen2: - benign (score < 0.15); + possibly damaging (score 0.15 - 0.85); ++ probably damaging (score > 0.85).

Table 8 *TMEM132D* variants located in ENCODE TFBSs.

TFBSs were identified using CHIP-Seq (chromatin immunoprecipitation with antibodies against the transcription factor and sequencing of the precipitated DNA). As TFBSs might overlap a genetic variant can be located in two or more sites. Adapted from Quast *et al.* [238].

Location SNV on chr12	Location SNV within gene	Location TFBS (Start)	Location TFBS (End)	TFBS for	Length TFBS
129564524	intron7	129564363	129564593	PU.1	194
129565100	intron7	129564992	129565336	GATA-2	333
129565175	intron7	129564992	129565336	GATA-2	333
129821986	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129821988	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129821990	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129821994	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129822017	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129822019	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129822051	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
129822129	intron4	129821939	129822223	STAT2	312
	intron4	129821896	129822186	STAT1	254
130015176	intron3	130015173	130015398	PU.1	889
	intron3	130015160	130015430	IRF4_(M-17)	292
130015328	intron3	130015173	130015398	PU.1	889
	intron3	130015160	130015430	IRF4_(M-17)	292
130015383	intron3	130015173	130015398	PU.1	889
	intron3	130015160	130015430	IRF4_(M-17)	292
130015620	exon3	130015615	130015975	STAT1	142
130015622	exon3	130015615	130015975	STAT1	142
130015810	intron2	130015615	130015975	STAT1	142
	intron2	130015636	130015912	Max	590
	intron2	130015641	130015877	FOSL2	233
	intron2	130015652	130015916	USF2	117
	intron2	130015687	130015958	STAT3	1000
	intron2	130015693	130015869	JunD	468
	intron2	130015710	130015814	USF-1	261
130385879	intron1	130385868	130386161	STAT3	143

SNV, single nucleotide variant; chr, chromosome; TFBS, transcription factor binding site

Location is according to the February 2009 Human Reference Sequence (UCSC Genome Browser).

3.1.5 Association analysis of variants in *TMEM132D* with AD

3.1.5.1 Association analysis of common variants

None of the tested common variants with a MAF higher than 5.0% (N = 32) showed an association with AD that survived correction for multiple testing. rs61945413 in intron 3 (p = 0.0016, OR = 0.65) and rs11060404 in intron 2 (p = 0.0017, OR = 0.66) showed a nominally significant association. While these two SNPs were in strong LD with each other (r^2 = 0.957), they were not in LD with the two common SNPs previously identified to be associated with PD (rs61945413: rs11060369 r^2 = 0.008 / rs7309727 r^2 = 0.019; rs11060404: rs11060369 r^2 = 0.01 / rs7309727 r^2 = 0.022).

3.1.5.2 Association analysis of rare and/or putatively functional relevant variants

None of the tested SNV sets showed a significantly different PMA in the *TMEM132D* locus between AD patients (N = 300) and controls (N = 300). In order to investigate patients with a more homogeneous phenotype, association analysis was restricted to patients suffering from PD with or without agoraphobia (N = 252). Although patients with specific or social phobias or GAD were excluded, significant associations could not be observed (Table 9).

Table 9 Association of the presence and the sum of rare and/or putatively functional alleles in the *TMEM132D* locus with AD and PD. Adapted from Quast *et al.* [238].

Marker set (N)	Presence of rare/functional alleles (N/%)			p-value[1]	
	Controls (N = 300)	Cases (N = 300)	PD patients (N = 252)	all cases controls	PD patients controls
MAF < 1% (N = 66)	57 (19.0)	53 (17.7)	46 (18.3)	0.678	0.800
coding (N = 25)	215 (71.7)	206 (68.7)	176 (69.8)	0.381	0.666
non-synonymous/splicing (N = 20)	180 (60.0)	165 (55.0)	138 (54.8)	0.202	0.209
Marker set (N)	**Mean number of rare/functional alleles (N/SD)**			**p-value[1]**	
	Controls (N = 300)	**Cases (N = 300)**	**PD patients (N = 252)**	**all cases controls**	**PD patients controls**
MAF < 1% (N = 66)	0.22 (0.480)	0.21 (0.505)	0.22 (0,519)	0.874	0.912
coding (N = 25)	1.39 (1.237)	1.25 (1.181)	1.22 (1.127)	0.167	0.090
non-synonymous/splicing (N = 20)	1.14 (1.227)	0.98 (1.161)	0.94 (1.106)	0.097	**0.044**

MAF, minor allele frequency; PD, panic disorder

[1] permutation p-value (1000 permutations), not corrected for multiple testing

For the SMA, no significant difference could be observed between AD patients and controls (Table 9). When testing only patients with PD, a significant association in the

SNV set, including non-synonymous and splicing variants could be identified (p empirical = 0.044). The SMA was higher in the control sample indicating a protective effect of this combination of variants. While PD patients carry more often none or one functional allele, individuals with more than two functional alleles are overrepresented in controls (Figure 11).

No significant association between any of the tested rare and/or putatively functional SNV sets and the previously identified common risk haplotype TA from rs7309727 and rs1106369 could be observed.

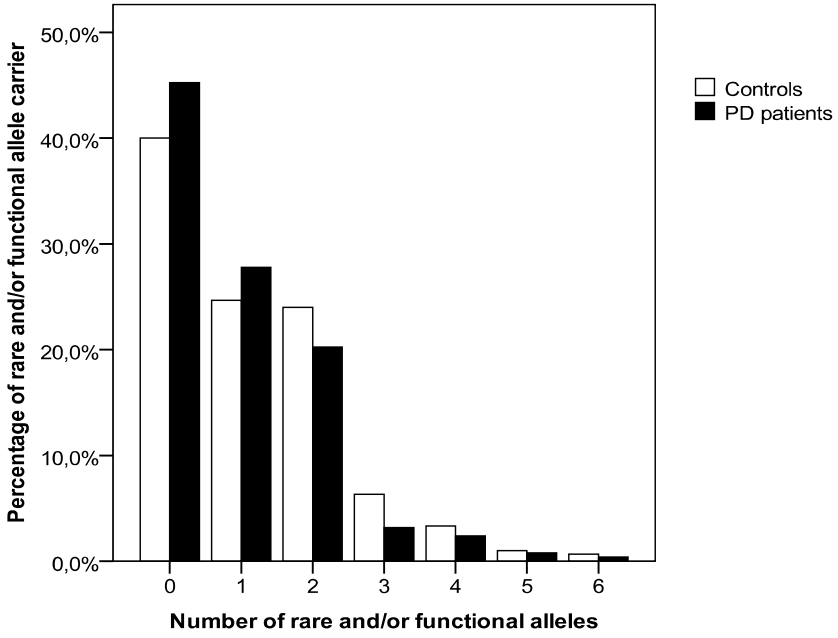


Figure 11 Distribution of rare and/or functional alleles in PD patients (N = 252) and controls (N = 300). The proportion of rare and/or functional allele carrier within each group is denoted on the y-axis. The shown data are based on the marker set containing non-synonymous and splicing variants (N = 20). Adapted from Quast *et al.* [238].

Besides the PMA and the SMA, the distribution of private coding variants between cases and controls was investigated. Private variants were defined as variants that only occur in either cases or controls of this study, but not in any other re-sequencing experiments such as ESP and 1000 Genomes Project. The rate of private variants was significantly increased in cases as nine patients compared to one control carried such variants ($0.01 < p < 0.05$, McNemar test). These nine patients had a nominally higher rate of relatives suffering from the same or any other form of AD than the remaining patients (66.6% versus 33.4%). Other phenotypic differences could not be observed (Table 10).

The most interesting private variant was chr12_128129066 which leads to an amino acid exchange from alanine to glycine in the extracellular domain of the TMEM132D protein. This variant was exclusively found in four unrelated patients of this study cohort, but not in any of the matched controls or in over 7,500 other individuals from the European, African and Asian population which were re-sequenced in the 1000 Genomes Project and the ESP.

Table 10 Clinical characteristics of AD patients with and without private coding variants. Adapted from Quast *et al.* [238].

Characteristics	Cases with private variants (N = 9)	Cases without private variants (N = 291)	p-value[1]
sex			
male	22.2% (2)	41.2% (120)	
female	77.8% (7)	58.8% (171)	0.319
HAM-A (SD)	18.3 (8.0)	24.1 (9.4)	0.069
HAM-D (SD)	13.2 (8.2)	13.8 (6.1)	0.784
PAS (SD)	26.1 (11.2)	30.0 (9.4)	0.258
age of onset (SD)	26.0 (9.2)	27.4 (11.5)	0.738
family history (any AD)			
yes	66.6% (6)	39.5% (115)	
no	33.4% (3)	60.5% (176)	0.165
additional psychiatric diagnosis			
yes	44.4% (4)	33.7% (98)	
no	55.6% (5)	66.3% (193)	0.495

HAM-A, Hamilton Anxiety Scale Score; HAM-D, Hamilton Depression Scale Score; PAS, Bandelow Panic and Agoraphobia Scale Score; AD, anxiety disorder [1] calculated using the Fisher exact test

3.1.5.3 Population stratification

In order to assess spurious associations due to population stratification, the method of genomic control was performed. A genomic inflation factor (λ) of 1.00594 was calculated. This implies that the associations observed in this study are indeed based on differences in case-control status and not on differences in allele frequencies due to different ethnic backgrounds. While λ did not suggest large effects of population stratification, a MDS plot identified three subjects outside from the main cluster of subjects (Figure 12). Therefore, these three subjects (two patients and one control) were excluded from the study sample and the association analyses involving rare variants were repeated.

However, all previous significant associations remained significant after removing these subjects.

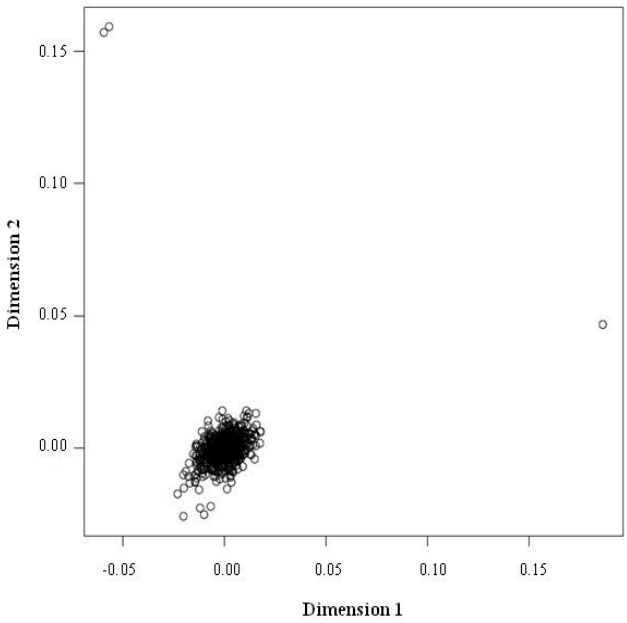


Figure 12 MDS plot based on genome-wide genotype data of 572 subjects.
Each data point indicates a sample. The first and second dimensions, which show the best segregation of the outliers, are represented. Taken from Quast *et al.* [238].

3.2 Results of the genetic and experimental functional characterization of the *SLC6A15* gene

3.2.1 Pooled targeted re-sequencing of the *SLC6A15* locus

Using the SOLiD 4 sequencer in order to re-sequence the whole *SLC6A15* gene, including 10 kb upstream and downstream (70 kb in total), in 400 MDD patients and 400 controls, 669 million raw reads with a length of 50 bp were generated in the first sequencing run (Table 11). 66.4% of these raw reads survived the QC and 85.1% of the quality filtered reads could be aligned to the reference sequence (NCBI Build 36.1, UCSC hg18). The average number of reads, which was included into the subsequent variant calling procedure, was thus 47 millions per DNA pool. The overall inclusion rate of the generated raw reads into variant calling was 56.2%.

Due to low coverage, two amplicons (12 kb in total) were re-sequenced in a second run using the SOLiD 5500xl machine. In this run, 253 million raw reads with 75 bp in length were generated. 89.3% of the raw reads had a sufficient quality for inclusion into the mapping step. From the QC filtered reads, 87.2% were mapped to the reference (GRCh37, UCSC hg19), resulting in approximately 24.6 million mappable reads per pool. Overall, 77.9% of the generated raw reads were mappable and could be used for further analyses.

Table 11 Number of reads obtained in the two SLC6A15 re-sequencing runs.
Reads were mapped using the BWA aligner. All reads are given in millions. Adapted from Quast *et al.* [242].

DNA Pool	NGS Run 1			NGS Run 2		
	raw reads	QC filtered reads	mappable reads	raw reads	QC filtered reads	mappable reads
Cases 1	84.6	57.6	51.0	26.6	23.8	20.7
Cases 2	83.6	56.0	46.8	33.1	29.6	25.7
Cases 3	84.0	56.3	48.0	26.6	23.9	20.9
Cases 4	90.8	59.0	49.3	42.9	38.4	33.6
Controls 1	79.4	52.6	45.4	28.3	25.2	21.7
Controls 2	96.4	62.6	54.1	23.0	20.6	17.8
Controls 3	64.6	43.1	36.0	38.1	34.0	29.8
Controls 4	85.3	56.7	47.2	34.1	30.5	26.5
Total	668.6	443.9	377.7	252.8	226.0	196.8

NGS, next-generation sequencing; QC, quality control

In the first NGS run, the average coverage was approximately 33,000 fold per base position per DNA pool and therefore theoretically 330 fold per subject (100 MDD patients

or controls per pool) and 165 fold per allele (200 alleles per pool). 81.4% of the whole sequenced region and 96.3% of the protein coding sequence were included into the SNV calling procedure due to a mean coverage > 5,000 at a given base position in each pool. With an average sequencing depth of 140,000 per base per DNA pool, 1,400 per individual and 700 per allele, the coverage in the second NGS run was approximately four times higher than in the first run. This can be explained by the fact that only 12 kb instead of 70 kb of the SLC6A15 locus were sequenced in the second run. A mean coverage above 5,000 in each pool was obtained for 74.3% of the sequenced bases and 99.2% of the bases in protein coding regions.

3.2.2 Genetic variants in the *SLC6A15* gene

In total, 405 genetic variants were detected in the two re-sequencing runs. Of these variants, 218 (53.8%) have not been previously published in the dbSNP137 database (accessed in April 2013). Furthermore, almost the same number of variants (N = 225, 55.6%) has not been identified in the 1000 Genomes Project database (April 2012 release). More than 50% (N = 225) of the detected SNVs were rare, with a MAF <= 0.5%, i.e. four or less occurrences among the 800 screened individuals. Of these extremely rare variants, 61.8% (N = 139) were novel and not already present in dbSN137. In contrast, from the 60 common variants with a MAF > 5.0% only one variant was not previously published in the dbSNP137 database.

Comparable with the TMEM123D re-sequencing study, the vast majority of the detected variants was located in non-coding regions of the gene. Three variants (0.7%) were in the 5‘ UTR, 44 (10.9%) in the 3’ UTR and 257 (63.4%) in intronic sequences. 85 variants (21.0%) were 5’ or 3’ of the gene locus. 16 variants (4.0%) were in the protein coding regions of the gene, twelve of those leading to amino acid exchanges in the protein. Of these tweleve non-synonymous variants, seven have been previously reported in dbSNP137 and eight have been identified in the ESP database (accessed in April 2013).

3.2.3 SLC6A15 variant validation using Sequenom re-genotyping

In the case-control MDD discovery sample (N = 800), 71.2% of the successfully re-genotyped SLC6A15 variants (N = 66) could be confirmed as polymorphic (Figure 13). The set of validated variants included nine non-synonymous variants, five of them only present in the long isoform of the gene (NM_182767) and three of them restricted to the shorter isoform (NM_018057) (Table 12). For the validated 47 variants, the correlation between the MAFs estimated from the NGS experiment and verified by individual re-genotyping was excellent with *r = 0.996.*

The nine non-synonymous variants, which were polymorphic in the discovery sample, were re-genotyped in the replication sample which consists of 905 MDD patients and 1029 controls. The set of variants for the replication stage was supplemented by 22 additional non-synonymous SNVs from the ESP database which have not been detected in the discovery sample (Table 13). Six of the non-synonymous variants detected in the NGS experiment were also polymorphic in the replication sample. However, for these six variants no consistent direction for overrepresentation in cases versus controls could be observed (Table 13). From the ESP variants, only three were polymorphic in the replication sample, two of them only in a single individual.

For the validated non-synonymous variants (N = 12), which were also present in the ESP database (N = 11), the correlation between the MAFs denoted in the ESP for the European American population and obtained from the re-genotyping experiment in either the discovery, the replication or the combined sample was excellent (*r = 0.999*).

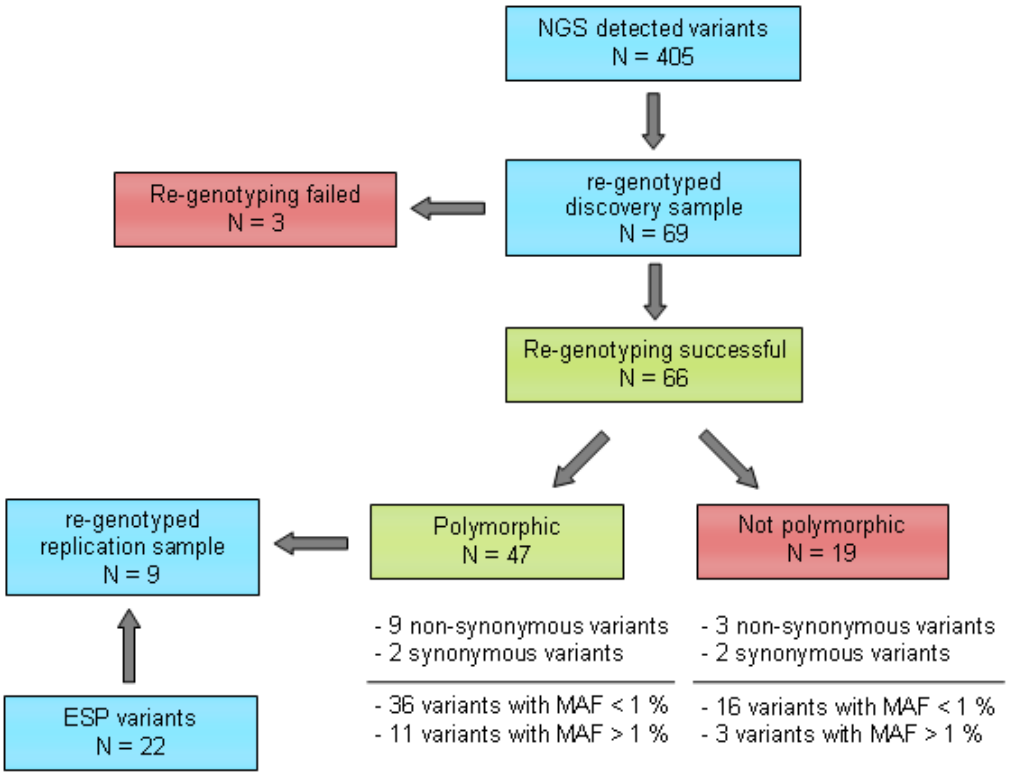


Figure 13 Validation of detected SLC6A15 variants performing Sequenom re-genotyping.
Non-synonymous variants which were polymorphic in the discovery sample and additional non-synonymous variants from the ESP database were re-genotyped in the replication sample. Denoted MAF was estimated from NGS discovery stage. Adapted from Quast *et al.* [242].

Table 12 Validated non-synonymous variants in the *SLC6A15* locus combined with potential functional effects assessed by *in silico* analysis. Adapted from Quast *et al.* [242].

				Validation stage								NT conservation		AA conservation		
SNV	Location on chr12	SNV in dbSNP137	Allele	initially found in	MAF Case (%)	MAF Con (%)	OR[1]	N[2]	mRNA isoform[3]	Location within gene	AA exchange	Splicing analysis[4]	Phast Cons[5]	phyloP[6]	SIFT[7]	Panther[8]	Poly Phen2[9]
chr12_83809886	85285755	rs139354471	T > C	NGS	0.08	0.11	0.7	2734	long/short	exon 2	T49A	no effect	-	+	-	+	-
chr12_85277713	85277713	rs200478124	C > A	ESP	0.06	np	Case	1934	long/short	exon 5	K227N	ESE site broken ESE site broken/new	+	+	-	+	+
chr12_83801746	85277615	rs79063785	A > G	NGS	0.15	0.21	0.7	2734	short	exon 5	L260P	ESS site	-	-	+	+	+
chr12_83801723	85277592	rs77477149	C > T	NGS	0.15	0.21	0.7	2734	short	exon 5	G268R	ESE site broken	-	-	-	-	-
chr12_83801692	85277561	rs17183577	T > A	NGS	17.50	19.25	0.9	800	short	exon 5	D278V	new ESE site	-	-	-	-	-
chr12_83790615	85266484	rs12424429	G > A	NGS	1.04	0.88	1.2	2734	long	exon 8	A400V	no effect	+	-	-	-	-
chr12_83790552	85266421	-	A > G	NGS	np	0.13	Con	800	long	exon 8	L421P	ESE site broken ESE site broken/new	+	+	+	-	-
chr12_83785100	85260969	rs201461650	A > G	NGS	0.08	0.07	1.1	2734	long	exon 10	I500T	ESS site	+	-	-	+	-
chr12_85257265	85257265	rs138060449	T > C	ESP	0.22	0.10	2.3	1934	long	exon 11	N591D	no effect	+	+	-	+	-
chr12_85257235	85257235	-	C > T	ESP	np	0.05	Con	1934	long	exon 11	A601T	new ESS site	+	+	+	+	++
chr12_83779683	85255552	rs145111717	C > A	NGS	0.31	0.21	1.5	2734	long	exon12	E684D	no effect	+	-	+	-	++
chr12_83779607	85255476	rs144267969	C > T	NGS	0.13	np	Case	800	long	exon 12	G710R	no effect	+	+	+	-	++

SNV, single nucleotide variant; chr, chromosome; NGS, next-generation sequencing; ESP, Exome Sequencing Project; MAF, minor allele frequency; Con, control; np, not polymorphic; OR, odds ratio; N, number of individuals; AA, amino acid; ESE, exonic splicing enhancer; ESS, exonic splicing silencer; NT, nucleotide

Location on chr12 is according to the February 2009 Human Reference Sequence (UCSC Genome Browser). Known SNVs are recorded in the dbSNP137 database.

[1] Case: SNV only in cases; Con: SNV only in controls; [2] N = 2734, variant was present in both the discovery and the replication sample; N = 800, variant was only polymorphic in the discovery sample; N = 1934, variant was only polymorphic in the replication sample [3] Long isoform is according to the RefSeq annotation NM_182767, the short isoform NM_018057 [4] FASTSNP; [5] ANNOVAR, phastCons 46-way alignment; [6] ANNOVAR, phyloP alignment, restricted to non-synonymous variants: + conserved (score > 0.95), - non-conserved (score < 0.95) [7] SIFT (sorting intolerant from tolerant): - tolerant (score > 0.05), + possibly damaging (score < 0.05) [8] Panther: - unlikely functional effect (pdeleterious < 0.5), + possibly damaging (pdeleterious > 0.5) [9] PolyPhen2: - benign (score < 0.15), + possibly damaging (score 0.15 - 0.85), ++ probably damaging (score > 0.85).

Table 13 Summary of the re-genotyping of all non-synonymous variants in the discovery sample, replication sample and combined sample. Adapted from Quast *et al.* [242].

SNV	Location SNV on chr12	initially found in	Discovery sample (N = 800)			Replication sample (N = 1934)			Combined sample (N = 2734)		
			MAF Case (%)	MAF Con (%)	OR	MAF Case (%)	MAF Con (%)	OR	MAF Case (%)	MAF Con (%)	OR
chr12_85285806	85285806	ESP	not re-genotyped			np	np				
chr12_83809886	85285755	NGS	0.13	0.13	1.0	0.06	0.10	0.6	0.08	0.11	0.7
chr12_85285676	85285676	ESP	not re-genotyped			np	np				
chr12_85279737	85279737	ESP	not re-genotyped			np	np				
chr12_85277713	85277713	ESP	not re-genotyped			0.06	np				
chr12_85277622	85277622	ESP	not re-genotyped			np	np				
chr12_83801746	85277615	NGS	0.25	0.13	2.0	0.11	0.24	0.5	0.15	0.21	0.7
chr12_83801723	85277592	NGS	0.25	0.13	2.0	0.11	0.24	0.5	0.15	0.21	0.7
chr12_85277576	85277576	ESP	not re-genotyped			np	np				
chr12_85277573	85277573	ESP	not re-genotyped			np	np				
chr12_83801692	85277561	NGS	17.50	19.25	0.9	assay failed					
chr12_85266930	85266930	ESP	not re-genotyped			np	np				
chr12_85266927	85266927	ESP	not re-genotyped			np	np				
chr12_85266902	85266902	ESP	not re-genotyped			np	np				
chr12_85266562	85266562	ESP	not re-genotyped			np	np				
chr12_83790615	85266484	NGS	1.75	0.63	2.8	0.72	0.98	0.7	1.04	0.88	1.2
chr12_85266469	85266469	ESP	not re-genotyped			np	np				
chr12_83790552	85266421	NGS	np	0.13	np	np	np				
chr12_85264301	85264301	ESP	not re-genotyped			np	np				
chr12_85264278	85264278	ESP	not re-genotyped			np	np				
chr12_85264267	85264267	ESP	not re-genotyped			np	np				
chr12_83785100	85260969	NGS	0.13	0.13	1.0	0.06	0.05	1.1	0.08	0.07	1.1
chr12_85260925	85260925	ESP	not re-genotyped			np	np				
chr12_85257357	85257357	ESP	not re-genotyped			np	np				
chr12_85257265	85257265	ESP	np	np		0.22	0.1	2.3			
chr12_85257235	85257235	ESP	not re-genotyped			np	0.05				
chr12_83779683	85255552	NGS	0.5	0.25	2.0	0.22	0.2	1.1	0.31	0.21	1.5
chr12_85255550	85255550	ESP	not re-genotyped			np	np				
chr12_85255544	85255544	ESP	not re-genotyped			np	np				
chr12_83779607	85255476	NGS	0.1	np		np	np				
chr12_85255472	85255472	ESP	not re-genotyped			np	np				

SNV, single nucleotide variant; chr, chromosome; NGS, next-generation sequencing; ESP, Exome Sequencing Project; MAF, minor allele frequency; OR, odds ratio

3.2.4 Case-control association analysis

As the selection of variants for validation was biased towards putatively functional relevant variants, which are mostly skewed to lower frequencies, only three common variants were individually re-genotyped. From these variants, only two were validated to be polymorphic in the discovery sample. Testing both variants for association with case-control status did not show any significant results. For the tested SNV set, including nine non-synonymous variants, no significant differences in the SMA and PMA between depressed patients (N = 1305) and controls (N = 1429) of the combined sample could be observed.

3.2.5 *In silico* functional annotation of non-coding variants in *SLC6A15*

Mapping all 405 detected variants to ENCODE TFBSs of different tissues, including five neuroblastoma cell lines, 15 intronic and three intergenic variants with potential influences on gene transcription were identified (Table 14). Interestingly, these 18 variants were also identified to overlap with DNaseI hypersensitivity sites in brain, including cerebellum, frontal cerebrum and frontal cortex. From these variants, one variant with an OR of 2 and three variants, which were also identified to be located in conserved regions of the genome (see below), were re-genotyped in the discovery sample. Two variants, which were previously reported in dbSNP137, could be validated. Variants, which disrupt putative miRNA target sites in the 3'UTR of genes, were not observed. Using PhastCons, seven non-coding variants in conserved regions of the genome were annotated. These variants were re-genotyped in the discovery sample and one variant upstream, one variant in intron 1 and two variants in the 3'UTR of the gene could be validated. While the intronic and the upstream variant were already reported in dbSNP137, the two 3' UTR variants were unknown so far. As the ORs of the validated variants were around 1, they were not re-genotyped in the replication cohort.

Table 14 *SLC6A15* variants located in ENCODE TFBSs which were identified using ChiP-Seq. Adapted from Quast *et al.* [242].

Location SNV on chr12	Location SNV within gene	Location TFBS (Start)	Location TFBS (End)	TFBS for	Length TFBS
85304592	intron 1	85304469	85304733	YY1_(C-20)	264
85304667	intron 1	85304469	85304733	YY1_(C-20)	264
	intron 1	85304649	85305085	Pol2	436
85304707	intron 1	85304469	85304733	YY1_(C-20)	264
	intron 1	85304649	85305085	Pol2	436
85304824	intron 1	85304649	85305085	Pol2	436
	intron 1	85304823	85305027	TAF7_(SQ-8)	204
85304851	intron 1	85304649	85305085	Pol2	436
	intron 1	85304823	85305027	TAF7_(SQ-8)	204
85304862	intron 1	85304649	85305085	Pol2	436
	intron 1	85304823	85305027	TAF7_(SQ-8)	204
85304863	intron 1	85304649	85305085	Pol2	436
	intron 1	85304823	85305027	TAF7_(SQ-8)	204
85304936	intron 1	85304649	85305085	Pol2	436
	intron 1	85304823	85305027	TAF7_(SQ-8)	204
85305066	intron 1	85304649	85305085	Pol2	436
	intron 1	85304978	85305278	TBP	300
85305115	intron 1	85304978	85305278	TBP	300
85305172	intron 1	85304978	85305278	TBP	300
	intron 1	85305117	85305387	TAF1	270
	intron 1	85305119	85305458	Pol2	339
	intron 1	85305126	85305382	NRSF	256
	intron 1	85305138	85305744	ZNF263	606
	intron 1	85305140	85305384	PRDM1_(Val90)	244
	intron 1	85305148	85305324	JunD	176
	intron 1	85305161	85305425	YY1_(C-20)	264
85305575	intron 1	85305138	85305744	ZNF263	606
85305903	intron 1	85305644	85305914	TAF1	270
	intron 1	85305677	85305981	Pol2	304
85306174	intron 1	85306100	85306333	ZNF263	233
85306191	intron 1	85306100	85306333	ZNF263	233
85306844	upstream	85306425	85306903	Pol2	478
85306884	upstream	85306425	85306903	Pol2	478
85306903	upstream	85306425	85306903	Pol2	478

SNV, single nucleotide variant; chr, chromosome; TFBS, transcription factor binding site

3.2.6 Translation of non-synonymous coding variants in *SLC6A15* into function

3.2.6.1 *In silico* functional annotation

The functional relevance of the nine non-synonymous NGS variants, which were present in either the discovery sample only or in the combined sample, and the three ESP variants, which were polymorphic in the replication cohort, was first investigated performing computational analyses. Splicing analysis using FastSNP predicted seven non-synonymous variants to create new ESE or ESS sites, or to disrupt already existing

splicing motifs (Table 12). For only two non-synonymous variants, deleterious effects on the function of the gene product were predicted using SIFT, PolyPhen2 and Panther. For all other variants, the three amino acid conservation tools showed inconsistent results. The evolutionary nucleotide conservation prediction tools PhastCons and PhyloP identified five non-synonymous variants to be located in evolutionary conserved regions of the genome (Table 12).

3.2.6.2 Experimental functional annotation

In order to assess the functional consequences of all nine non-synonymous variants in the long human *SLC6A15* isoform, a proline uptake experiment was performed in HEK cells (see Table 5 in the methods section 2.6.2.1). The IC_{50} values for ^{3}H proline uptake did not differ between HEK cells transfected with plasmids containing the wild type (WT) *SLC6A15* sequence, and cells transfected with plasmids harbouring one of the nine non-synonymous variants in the *SLC6A15* gene (Figure 14).

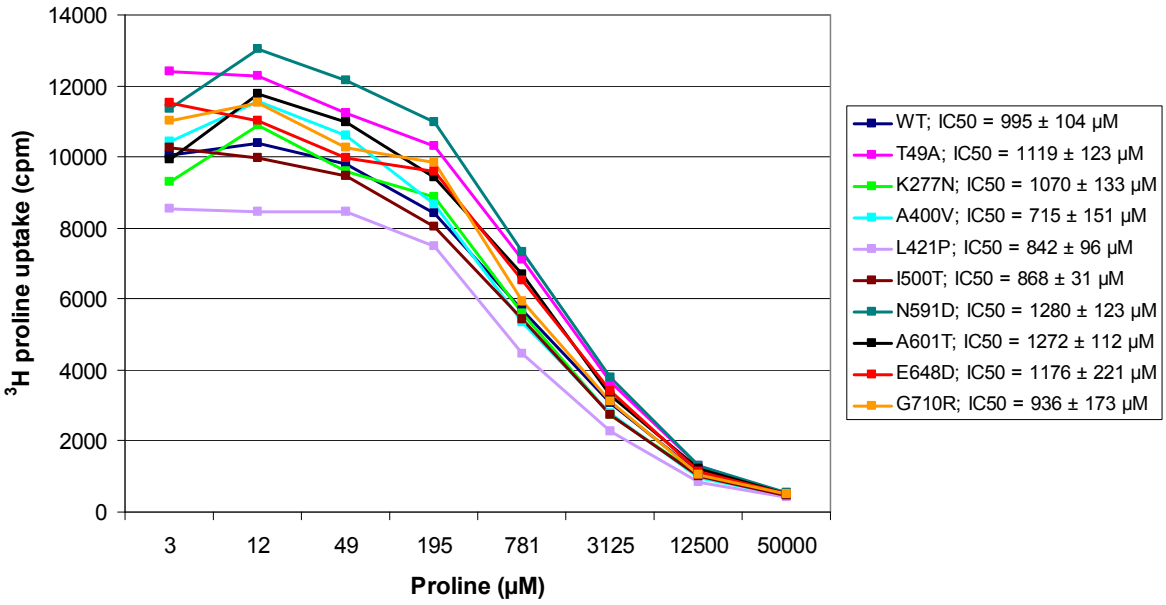


Figure 14 Inhibition of ^{3}H proline transport by the non-radioactive labeled amino acid L-proline. Concentration of non-labeled L-proline is plotted on the x-axis, ^{3}H proline uptake as counts per minute (cpm) on the y-axis. Each datapoint represents the mean transport activity of triplicate samples. Taken from Quast *et al.* [242].

While the IC_{50} values were not affected by non-synonymous variants in the *SLC6A15* gene, the maximal uptake of ^{3}H proline showed large differences, ranging from approximately 8600 to 12400 cpm (Figure 15). In order to confirm these findings, the three mutants with the largest alterations in maximal ^{3}H proline uptake (T49A, A400V

and L421P mutants) compared to HEK cells harbouring the WT plasmid were re-tested in a second independent uptake experiment.

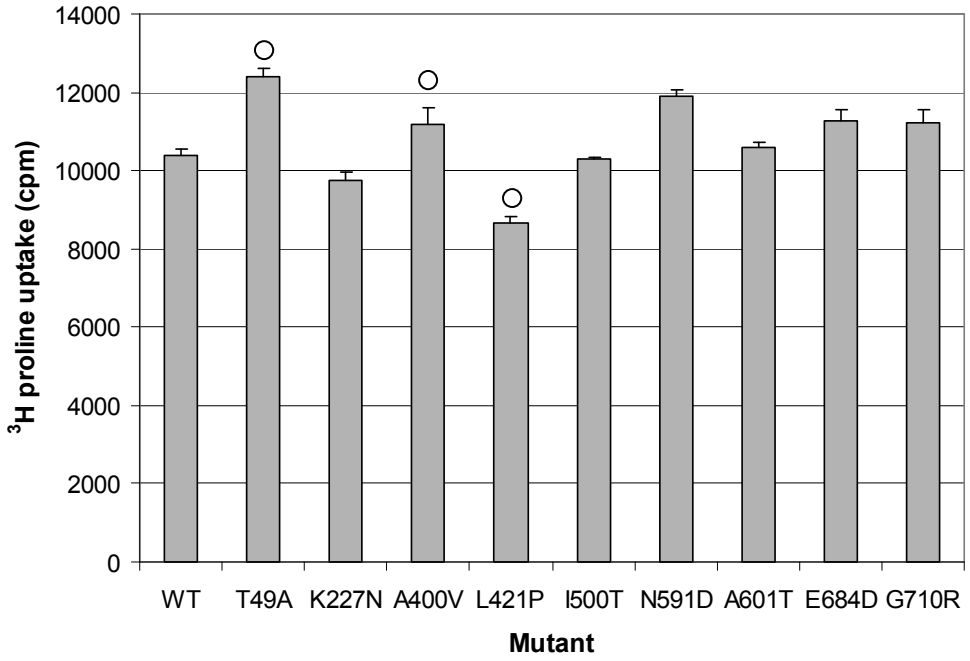


Figure 15 Maximum in ^{3}H proline uptake of WT and mutant HEK cells.
The maximum in uptake was measured in the presence of 3 µM non-labeled L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment. Taken from Quast *et al.* [242].

In the repeated uptake measurement, the results obtained in the first experiment could be replicated for all tested mutants (Figure 16). Significant differences in ^{3}H proline uptake could be observed across all concentrations (3 µM, 12 µM and 780 µM) of the non-labelled L-proline ($p = 1.8e\text{-}7$, in a two way ANOVA with mutant and non-labeled proline concentration as the two predictors and transfection efficiency as covariate ($F = 18.9$, $df = 2$). Mutant T49A and mutant A400V showed a significantly increased ^{3}H proline uptake compared to the WT, withstanding correction for multiple testing using the Bonferroni method ($p = 8.4e\text{-}9$ and $p = 0.001$ respectively). While mutant L421P showed a decrease in ^{3}H proline uptake as in the first experiment, this result was not significant after adjustment for multiple comparisons (p nominal = 0.016, p corrected = 0.158).

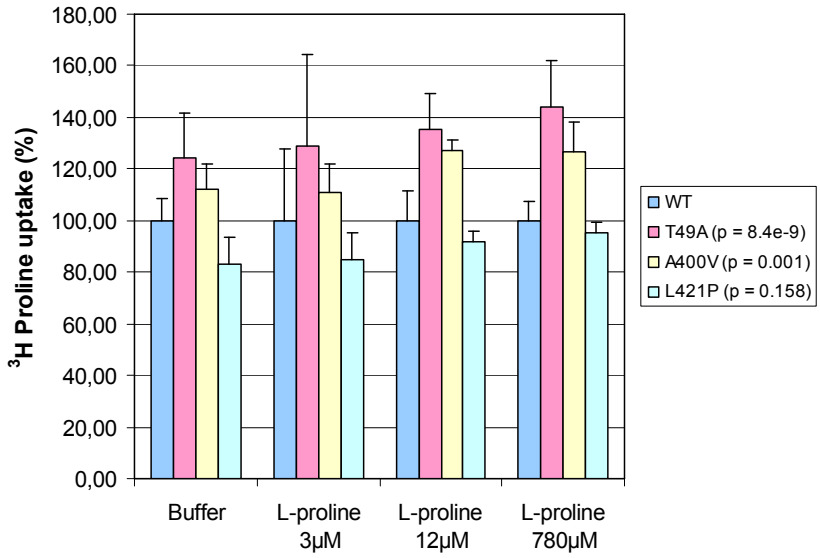


Figure 16 Repeated uptake measurement of mutants with large differences in maximal ^{3}H proline uptake compared to WT.

The uptake was measured under four different experimental conditions. Each bar represents the ^{3}H proline uptake (mean ± SD) obtained from triplicates for the buffer solution, six samples for the 3µM and 12 µM L-proline solution respectively, and nine samples for the 780 µM L-proline solution. Bonferroni corrected p-values, which are given in brackets, are based on the difference in mean ^{3}H proline uptake between WT and tested mutant across all concentrations of non-labelled L-proline. Taken from Quast *et al.* [242].

Fluorescence microscopy indicated that the sub-cellular localization of the SLC6A15 transporter to the cell membrane was not changed in any of the mutants harbouring a non-synonymous variant in the *SLC6A15* gene (Figure 17). In addition, these imaging experiments did not show any alterations in the level of the transporter at the cell membrane, indicating similar SLC6A15 expression levels in WT and mutant HEK cells.

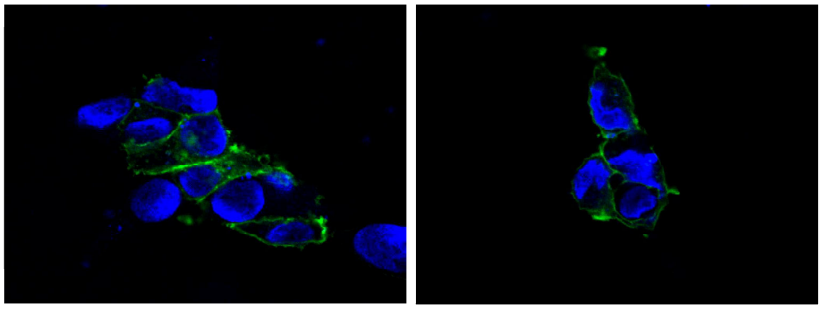


Figure 17 Sub-cellular localization of the SLC6A15 protein in WT (left) and T49A mutant cells (right).

The localization of the eGFP-h*SLC6A15* fusion product is indicated in green. Cell nuclei are stained with DAPI (blue). For all other mutants, fluorescence imaging showed similar expression patterns. Taken from Quast *et al.* [242].

4. DISCUSSION

4.1 Role of common and rare variants in the susceptibility to complex diseases

For more than a century, genetic epidemiology studies the question to which extent genetic variation contributes to the expression of a phenotype. GWAS have been successful in identifying thousands of common variants associated with complex diseases [2]. Despite this success, the majority of genetic variants contributing to complex traits have still to be discovered, as only a small proportion of the estimated heritability can be explained by these common variants [3]. The case of missing heritability has been the major motivation for the investigation of rare variants. Since the development of novel high-throughput sequencing technologies, which allow the identification of variants across the whole allelic frequency spectrum, many susceptibility genes for complex diseases have been screened for rare variants.

4.1.1 Importance of common and rare variants in disease susceptibility

While common variants have long been assumed to be the major factor for the susceptibility to develop common diseases, several arguments in favour for an essential role of rare variants were previously discussed. Besides evidence from evolutionary theory (see section 1.1.5.1 of the introduction), population genetic data further support the crucial role of rare variants in disease susceptibility. It has been shown that the distribution of genetic variants along the frequency spectrum is skewed towards rare variants, with over one third having MAFs below 5% [265]. In addition, the number of putatively functional relevant variants also increases with decreased MAF. Non-synonymous variants were identified to be significantly skewed towards low frequencies, while presumably non-functional variants segregate at higher frequencies, reflecting the purification selection of deleterious variants [266,267]. In the TMEM132D study, only 24.0% of the detected variants had MAFs above 5%. From the 13 non-synonymous variants, only one had a MAF higher than 5%. Similar, from twelve non-synonymous variants in the *SLC6A15* gene, only one was common, with a MAF of about 20.0%, and only 15.0% of all detected SLC6A15 variants were common. These data confirm the skew of variants towards lower frequencies and the enrichment of putatively functional variants in low frequency ranges so that the investigation of rare variants in order to uncover novel disease causal variants is warranted.

Another argument in favour of the importance of rare variants in common complex diseases comes from family studies. It has been shown that many rare Mendelian disorders, which accumulate in families, are caused by highly penetrant rare alleles with large effects. Several recent studies have been demonstrated that numerous complex disorders also have Mendelian subtypes in which rare variants with large effects cause the phenotype. Examples include rare variants promoting atherosclerosis through hypercholesterolemia [268], rare variants in the *BRCA1* and *BRCA2* breast cancer genes [269] and rare coding variants which are responsible for 25.0% of the cases with X-chromosomal linked intellectual disability [270].

The important role of rare variants in the susceptibility to common complex diseases could be further supported in the TMEM132D study. An increased number of private non-synonymous variants in AD patients compared to healthy controls could be observed (nine patients versus one control). The most interesting variant, which leads to an alanine to glycine exchange in the protein, was present in four non-related AD patients of the sample, but in none of the other samples, including samples of different ethnic origin. Interestingly, these patients had a nominally higher rate of family members with the same or another form of AD (66.6 versus 33.4%) than patients without private non-synonymous variants. The fact that these private variants are so rare (present in 1 – 4 individuals only) leads to the assumption that they are too new to be selected against due to deleterious effects on fitness. Therefore, extremely rare variants are highly likely to be functional and phenotypically relevant [271].

Besides an increased rate of private non-synonymous variants in AD patients, an overrepresentation of non-synonymous variants and variants with predicted changes on splicing in healthy controls as compared to PD patients was identified. In contrast to private variants, which are per definition extremely rare, these putatively functional relevant variants were distributed along a broad MAF spectrum, ranging from 0.17 to 30.0%. Combined with the data from the previous GWAS study, in which two common intronic variants were identified to be associated with PD [4], this pooled re-sequencing study suggests that not only common or rare variants alone, but a combination of both contributes to the development of anxiety-related phenotypes. Other studies in medical and psychiatric disorders have also begun to show that a combination of common and rare variants contributes to the susceptibility to common diseases [230-232].

The increased presence of common and rare functional variants in healthy controls leads to the suggestion that the combination of these presumably functional relevant variants in *TMEM132D* has a protective effect on PD. Variants leading to the protection against a

disease have also been identified for other disorders. Multiple rare coding variants were observed to contribute to low triglyceride and High Density Lipoprotein (HDL) cholesterol plasma levels which are associated with a protection against coronary atherosclerosis [272,273]. A decreased risk for type 1 diabetes was associated with four rare functional variants in *IFIH1* [274] and low IF1H1 levels have been found to be protective against this disorder [275].

In contrast to the overrepresentation of functional variants in controls, the increased rate of private non-synonymous variants in AD patients indicates that these variants have deleterious effects and increase thus the risk to develop AD. Hence, variants within the same gene can confer both increased risk as well as protection against a disease. In line with this suggestion, rare variants within the *PCSK9* gene were identified to be associated with higher levels of Low Density Lipoprotein (LDL) cholesterol, while others are associated with lower levels of LDL cholesterol [276].

In conclusion, the TMEM132D study demonstrates that both common and rare genetic variants contribute to the risk to develop common complex diseases. Hence, perhaps both of the long debated CDRV and CDCV hypotheses might be correct in some aspect. Although multiple rare variants were shown to play an important role in disease susceptibility, high effect sizes, which were postulated by the CDRV hypothesis, were not observed. Newer data from exome re-sequencing projects also indicated that rare variants do not contribute to disease risk with much higher ORs than common variants [277].

4.1.2 Additional factors contributing to the susceptibility to complex diseases

Although genetic variants play an important role in disease susceptibility, they are not the only factors which contribute to the manifestation of a disease. Gene-environment interactions, epigenetic factors and gene-gene interactions are also suggested to contribute to susceptibility to disease. Although these factors were not investigated in this thesis, a short description will be given in the following sections.

4.1.2.1 Gene-environment interactions in disease susceptibility

Several papers have begun to show that not distinct genetic or environmental causes alone are involved in the aetiology of physical and mental diseases, but an interaction between the two [13,135,136,278]. Hidden gene-environment interactions might be one possible explanation for the lack of significant associations of rare and/or common variants in *SLC6A15* with MDD. It might be possible that the detected variants are per se

not deleterious by themselves, but distinct environmental exposures, including sexual, physical or emotional abuse in childhood, premature parental loss or exposure to family conflict might be deleterious. Indeed, it has been shown that the presence of a depression susceptibility gene is not sufficient to develop the disease in the absence of environmental stressors [279]. In the reverse case, the occurrence of a severe traumatic event has little effect in the absence of a genetic susceptibility background, although stressful life events are among the strongest predictors of depression [279].

4.1.2.2 Epigenetic influences on disease susceptibility

Besides environmental influences, epigenetic factors, including DNA methylation and acetylation, also play a role in susceptibility to disease. The influence of methylation status on disease risk has been demonstrated by Klengel *et al.* among others [280]. In their study, a functional variant in the FK506 binding protein 5 gene (*FKBP5*), which is important for the regulation of the stress hormone system, was shown to increase the risk for stress-related psychiatric disorders in adulthood, when glucocorticoid response elements of *FKBP5* were de-methylated [281]. Interestingly, de-methylation could only be observed in individuals who experienced traumatic events in childhood. This is an impressive example for the importance of early environment for the later life. Early in the development of an individual, the expression of specific genes can be conditioned in a tissue specific manner [282,283]. It has been assumed that this developmental programming is based on early experiences in order to prepare an individual for the life under the experienced conditions [284]. Thus, negative experiences in early life are likely to predict later adversity in life. This programming of gene expression is maintained through the whole life by epigenetic modifications of the DNA and chromatin [282,285]. In the study of Klengel *et al.*, de-methylation of glucocorticoid response elements in the *FKBP5* gene might thus be caused by traumatic events in early life. While early environmental factors were already assumed to be involved in gene-environment interactions [279], the FKBP5 study suggests a possible mechanism how genes and environment interact and thus contribute to depression and other stress-related phenotypes.

4.1.2.3 Contribution of gene-gene interactions to complex diseases

Another form of multifactorial contribution to complex diseases, and a possible explanation for the lack of significantly associated variants in the *SLC6A15* locus with MDD, is gene-gene interaction. Evidence for epistatic effects and its implications for

depressive disorders came among others from a linkage study which has been demonstrated that epistasis between *SLC6A4* and a so far unknown gene on chromosome 4 is a risk factor for MDD [286]. In a neuroimaging study involving healthy subjects, an interaction between the 5-HTTLPR polymorphism and the *BDNF* Val66Met polymorphism could be observed [287]. The methionin allele of the *BDNF* gene was identified to be protective against the adverse effects of the short allele of the *SLC6A4* gene which has been reported to increase the risk for depression [288]. In contrast, the *BDNF* wild type allele was identified to support the depressiogenic effects of the short form of the *SLC6A4* locus. In a subsequent study, the *BDNF-SLC6A14* interaction could only be observed in individuals with childhood abuse, indicating a relevant impact of gene-gene-environment interactions on complex diseases [289]. In another study, the effects of child abuse on depressive symptomes were observed to be moderated by the interaction of genetic variants in the *CRHR1* locus and the 5-HTTLPR polymorphism [290].

4.2 Practical and statistical challenges of the novel NGS technologies

Compared to traditional Sanger sequencing, NGS technologies have an increased sample throughput which has led to dramatically reduced sequencing costs [221,291]. Due to this reduction in sequencing costs and the variety of possible applications, the NGS technique was selected by Nature Methods as the method of the year in 2007 [291]. Nevertheless, several major challenges of these novel NGS technologies exist.

4.2.1 Indications for and challenges of pooling approaches

Although NGS technologies have led to reduced sequencing costs, these costs are only one proportion of the overall costs which occur within the scope of a sequencing experiment. In addition, costs for DNA extraction, enrichment and preparation for sequencing have to be included into each cost calculation. Especially the costs for DNA preparation should not be neglected as, depending on the number of samples, these costs can easily exceed the sequencing costs. For instance, for a SOLiD barcoded fragment library preparation, costs of about 150 € per library can be estimated. In contrast, the costs for the SOLiD ToP sequencing chemistry account for about 1,200 € per run. Given that the library preparation in the SLC6A15 project would have to be performed for each of the 400 MDD patients and 400 controls individually, a budget of 120,000 € would be required solely for this first section of sample preparation.

The use of pooled DNA has been suggested to be an attractive cost-effective method to identify genetic variants in targeted re-sequencing approaches [292]. Here, DNA from different individuals is quantified, and equimolar amounts of the measured DNAs are mixed together. Especially for this type of NGS design, which is aimed on the detection of as much genetic variants as possible, pooled samples are highly recommended as the number of detectable variants increases with the number of samples. In contrast, whole genome or exome sequencing studies have mostly relatively small sample sizes so that the costs for sequencing are probably higher than the costs for sample preparation. In this case, individual sequencing may be more appropriate than DNA pooling.

A major drawback of DNA pooling is that the information, which genetic variant is present in which individual, is missing although this information is often required for further analyses. Therefore, DNA pooling is used as initial screening tool in the context of a two-stage design. In the first stage, DNA pools are sequenced in order to discover, in theory, all genetic variants within the target region. In the second stage, a subset of the identified variants, including for instance potentially functional relevant variants, is re-genotyped in all individuals of the discovery sample using an independent method.

The question, which often arises in the context of pooling approaches, is how many individuals to combine in a pool. The extraction of as much statistical information as possible, at cost and work load as low as possible, is the most important factor for determining the pool size. Statistical information can be defined as the power to detect a genetic variant in a pool. When the size of a DNA pool increases, the number of alleles also increases which leads to a higher probability to detect a variant. However, the frequency of a variant decreases with increased pool size, so that rare variants might fall under the detection threshold which decreases the variant detection probability [293]. The detection threshold, which is given by the sequencing error rate of the sequencer, is the major factor that limits the size of DNA pools in which a single heterozygous allele remains detectable. Given that a pool of 50 individuals is sequenced with a sequencing error rate of 1%, one can not decide any more whether one altered allele is due to a true variant or a sequencing error at that position. A high error rate can be specifically critical for analysing pooled samples, because sequence analyses of pools derived from a large number of individuals is prone to erroneous variant calling [294]. To address the problem of erroneous variant detection, a minimum in coverage at a given base position in each of the sequenced pools for inclusion of a base into variant calling procedure, is recommended.

4.2.2 Uneven coverage distribution and its implication for variant discovery

The base composition of the human genome has already been a challenge to Sanger sequencing and it has continued to be a major problem for NGS methods [295]. It has been demonstrated that DNA amplification performing PCR is biased towards regions with balanced base composition and that less complex regions with high AT or GC content are less or not amplified [296,297]. This bias, which occurs during library amplification, leads to an uneven coverage as the generated reads are not uniformly distributed along the sequenced region. In the TMEM132D study, the mean coverage across the whole sequenced region was about 50,000 fold per base per pool. However, the actual coverage at a given base position within each pool varied dramatically, ranging from 300 to 400,000 fold. In contrast, the overrepresentation of reads at the ends of the amplicons, compared to the middle of the amplified DNA fragments, is not a sequence specific problem, but results from DNA fragmentation where nucleotides located at amplicon ends are fragmented more frequently than nucleotides in the middle [298,299].

The large fluctuations in sequencing depth demonstrate its implication for variant discovery. As already described in section 4.2.1, a minimum in coverage is required for inclusion of a base into variant calling. Uneven coverage distribution increases the likelihood that bases are insufficient covered by reads and thus excluded from subsequent variant analyses, although these positions might harbour genetic variants. In the TMEM132D project, about 10.0% of all sequenced bases had a coverage below the required minimum of 5,000 in each pool. However, none of the bases in the coding region of the *TMEM132D* locus was excluded. This might be explained by the fact that the sequence complexity of introns is expected to be reduced compared to exons since more repetitive elements are present in non-coding regions [300]. Further evidence for a sequence specific bias in coverage came from the comparison of the coverage patterns of the DNA pools. While the sequencing depth varied dramatically within each pool, the coverage distribution was highly conserved across all four TMEM132D pools.

To overcome the problem of uneven coverage, it is recommended to use as much starting material as possible for library preparation in order to avoid the amplification step. If PCR amplification can not be avoided, it is suggested to keep the number of PCR cycles as low as possible.

4.2.3 Alignment of short sequencing reads as a statistical challenge

The most fundamental step for almost all NGS applications is the mapping of sequencing reads to the reference genome [301]. Alignment, which is the finding of the most credible source for the sequenced DNA fragment, is challenging due to the length of the generated sequencing reads. While the Sanger-based sequencing method provides reads with up to 900 bp, sequences, which are provided by a NGS sequencer, are much shorter, ranging from 30-700 bp. The length of a DNA sequence is a crucial factor for the uniquely alignment. The shorter a read, the higher the probability that the sequence will align equal to multiple chromosomal locations. Especially reads, which derive from regions with low complexity, map not only to the targeted region, but also to a high extent to genomic regions outside the investigated one. This alignment bias towards regions with higher complexity is another reason for an uneven coverage distribution across the sequenced region.

Mapping tools, which are commonly used for Sanger sequencing, are not adapted for NGS purposes as these tools are not designed to align reads with 700 bp at maximum. To meet this challenge, many mapping tools have been introduced flooding the market until now [302]. In this thesis, read alignment was performed using BWA [252] and SHRiMP [253]. Consistent with the results of a study in which several aligners were compared with each other, BWA clearly outperformed SHRiMP regarding mapping speed and memory capacity [303]. In this comparison study, BWA was shown to require approximately ten times less memory occupancy than SHRiMP for mapping, and processes the data approximately 30 times faster than SHRiMP. These disadvantages of the SHRiMP aligner can be outbalanced by a higher mapping sensitivity and accuracy. Both the percentage of reads, which can be mapped to the reference (sensitivity), and the percentage of reads, which were mapped correctly (accuracy), are increased when using the SHRiMP aligner [303]. After SHRiMP alignment, about 600 variants within the sequenced *TMEM132D* region were discovered, while only about 400 variants were called after mapping the reads using BWA, confirming the higher sensitivity of SHRiMP. An increased number of called variants due to a higher mapping rate could also be observed in the SLC6A15 project (600 called variants using SHRiMP versus 500 variants using BWA). Increased accuracy of the SHRiMP aligner was shown by comparing the validation rate from a subset of called variants which was individually re-genotyped on the Sequenom platform. 95.0% of the SHRiMP aligned variants in *TMEM132D* could be validated, while only 83.0% could be verified in the re-genotyping experiment as true variant when they were called after BWA alignment, indicating a lower false-positive rate

of SHRiMP. On the other hand, the percentage of falsely not called variants was increased in SHRiMP as it did not detect 21 already validated variants. Finally, despite the increased rate of false positives, the BWA aligner was used for the TMEM132D project as this disadvantage was accepted for a lower false negative rate, higher mapping speed and lower memory capacity. For the SLC6A15 project, both aligners were used and only variants, which were called in both approaches, were included for further analyses in order to minimize the false discovery rate.

4.3 Rare genetic variants as a challenge for genetic association studies

Since the establishment of high throughput sequencing methods, the identification of genetic variants, even those with a private character, has become a standard method in human genetics. While the detection of rare variants has become increasingly easy, association analysis and sample recruitment remain difficult.

4.3.1 Association testing of rare variants as a statistical challenge

Statistical methods for the detection of associations of common variants have been extensively developed and successfully applied to numerous studies of complex diseases. These methods are based on single-marker tests, whereby an individual marker is tested for an association with disease using univariate statistical tests. Unfortunately, most of these single-marker testing tools are unsuitable for association analysis involving rare variants, as the power to detect an association with a single rare variant is low, even in very large samples [97,257]. In general, the sample size, which is required to detect an associated variant, increases linearly with 1/MAF [83]. With 300 AD patients and 300 controls, the sample of the TMEM132D study was only sufficiently powered (power > 0.8) to detect associations of variants with a MAF of 1% and an OR of 3.5 (additive model, uncorrected alpha level 0.05). To detect the non-synonymous variant T49A (MAF 0.1%, OR 0.7), which was identified to significantly increase the activity of the SLC6A15 amino acid transporter, with a power of 0.8 at an alpha level of 0.05, a case-control sample of over 70,000 would be required. Both examples highlight the difficulty to detect associations of rare variants.

To overcome the problem of low power, several alternative association approaches have been developed, suggesting the assessment of the collective effects of multiple rare variants within and across genomic regions [257,304]. This collapsing method, which is also referred to as burden testing, counts individuals who carry one of the variants of a

marker set, calculates the frequencies of these individuals in different groups such as cases and controls, and tests then the two groups for frequency differences.

Different approaches, which variants to include into a marker set and to test this combination of variants for association, exist. First of all, variants might be selected based on their location within the genome. Second, different frequency thresholds can be used for selection. Third, rare variants can be grouped according to their functionality. Combining variants according to their functional consequenes is a very popular approach as it has been demonstrated that testing a marker set, which comprises possible functional relevant variants, is highly advantageous for the identification of disease susceptibility variants [223,305]. In this thesis, a significant association between genetic variants in *TMEM132D* and PD could only be observed, when testing the marker set, including non-synonymous variants and variants with predicted effects on splicing. Collapsing variants with a MAF below 1%, irrespective of possible functional effects, did not show any significant associations. Similar results have been reported by Davis *et al.* who investigated the association of variants in the *TTC21B* locus with human ciliopathies [306]. An overrepresentation of rare alleles in controls compared to cases could only be observed when restricting the coding variants to those with predicted function. This study also demonstrates the advantage of testing functional relevant variants in order to detect disease susceptibility variants. Information about putatively functional effects can be gained by different approaches which are described in section 4.4.

Several statistical analysis tools can be used to test the hypothesis that a combination of variants is associated with a disease. The simplest approach, which was also used in this work, is the Cohort Allelic Sum Test (CAST) which compares the number of individuals who carry one or more variants of the tested marker set between case and control group. The Combined Multivariate and Collapsing (CMC) method is an extension of the CAST method [257]. In this method, all rare variants with a MAF < 1% are collapsed, and the collapsed variants are treated as a single common variant which is then analysed together with other common variants in a multivariate analysis. A combined analysis of common and rare variants is recommended as a number of studies have been shown that variants within a wide range of frequencies are involved in disease aetiology [307-309]. For example, for HDL cholesterol, common and rare variants were detected to have modifying effects on HDL cholesterol levels [310]. Another extension of the original CAST method is the weighted sum collapsing approach [311]. In this approach, rare variants are given more weight because stronger effects are expected for rare variants than for more frequent variants.

While none of the described approaches differentiate between protective, neutral or deleterious variants, Han and Pan developed a method which considers the direction of the effects of the variants [312]. This approach is highly recommended as all above mentioned methods assume that the effects of all variants, which are included into a collapsed marker set, have the same direction. However, this assumption is rather unlikely, unless the number of variants is low. For instance, it has been shown that rare variants within the *PCSK9* gene are associated with higher levels of LDL cholesterol, while others in the same gene are associated with lower levels of LDL cholesterol [276].

4.3.2 Rare variants as a challenge for the study design

The recruitment of individuals for an association study involving rare variants is a crucial issue, as rare variants are more likely to be population specific than common variants so that allele frequencies might vary drastically between different populations, independent of disease status [92,313,314]. Hence, differences in ethnic background can lead to spurious associations [315,316]. The problem of population stratification in association analyses occurs both in case-control studies and in studies, examining individuals from the extreme end of a phenotype [273,276,317], while the investigation of related individuals is not affected by spurious associations due to population stratification.

In order to avoid spurious associations due to population stratification, carefully performed matching of cases and controls is essential. In general, the population, from which the controls are derived, should be the same, from which the cases came from. Similarly, the population, from which the individuals with extremes in phenotype are derived, should be the same population, in which detected variants are planned to be validated. Recruitment of individuals from the same geographical region, or self-reports regarding family ancestry are attempts to keep the level of population stratification low. Nevertheless, it is impossible to match for all genetic differences and thus, statistical methods are required to detect population stratification.

In the TMEM132D study, the method of genomic control identified a genomic inflation factor (λ) of 1.00594 which implies no effects of population stratification. Values of $\lambda < 1.05$ are generally considered to be benign [318]. If λ is above 1.05, all association test statistics should be corrected for background population stratification by the genomic inflation factor [259]. The method of PCA identified three individuals (two patients and one control) which were outside of the main cluster of individuals. However, after excluding these three subjects and repeating asociation testing, the previous observed significant association remained significant. Thus, the overrepresentation of putatively

functional relevant variants in controls compared to PD patients was indeed associated with case-control status and not with differences in allele frequencies due to different ethnic backgrounds.

4.4 Functional characterization of genetic variants in association studies

While thousands of significant associations have been turned up in GWASs in the last years, only very few of the variants identified to be correlated with a disease have been shown to be the actual risk variant [319]. This can be explained by the fact, that association does not mean causation and that association alone does not imply functional relevance. Therefore, functional characterization of genetic variants is highly recommended in association studies, in order to check whether the associated variant itself or another variant in LD is responsible for the investigated phenotype. Furthermore, burden testing has been shown to be more powerful when marker sets, containing functional relevant variants with the same direction of effect, are tested for association.

While the functional annotation of genetic variants is of paramount interest, the accomplishment of this task is challenging. A human individual is estimated to carry about 3.7 million SNVs distributed across the whole genome [35]. The number of variants in the coding regions of the genome is estimated to 20,000 – 24,000, including 10,000 – 11,000 non-synonymous variants that could negatively, but also positively influence the function of a gene [95]. Indeed, coding variants, resulting in amino acid substitutions, premature stop sites or deleted parts of a gene, are heavily enriched among disease causing variation in Mendelian disorders [8]. Variants in the non-coding genome, which accounts for 99% of the total human genome, might also be of functional relevance. Variants in regulatory regions such as TFBSs, enhancers and promoters have been shown to predominate as risk factors for common disorders [320]. The large number of putatively functional relevant variants makes functional characterization laborious and the identification of the causal variant for a disease often comparable with the finding of a needle in a haystack.

4.4.1 Computational functional annotation

Computational approaches provide an easy and fast possibility to predict whether a genetic variant is functional relevant or not. In the last years, a number of tools, which are based on the principle of sequence homology between organisms, have been developed [321]. It is assumed that deleterious variants are more likely at positions of the genome that are evolutionary conserved and have not been removed by natural

selection [247,322]. One major disadvantage of these prediction methods is that results obtained from different tools can not be compared directly. Direct comparisons are problematic as each tool uses different algorithms and sequence databases as reference for their deleteriousness estimation. In addition, some tools include information about the structure of the protein into their predictions [323,324]. Variants in the interior of a protein are suggested to have larger effects on protein function than variants at the outer side of the gene product. Other tools additionally incorporate biochemical data such as positions of active sites and disulfide bridges or charge of amino acids [325]. The integration of structural and biochemical information to comparative sequence analysis is suggested to significantly improve predictions of deleteriousness [326,327].

In this thesis, the three protein-sequence based prediction tools SIFT, PolyPhen2 and Panther were used. From twelve validated non-synonymous variants in the *SLC6A15* locus, only two were predicted to be deleterious in all three tools. For all other variants, the interpretation was difficult due to inconsistent predictions. A similar picture could be observed, when subjecting all 13 non-synonymous variants in *TMEM132D* to computational functional annotation, as none of the variants was predicted to have a deleterious effect in all three tools. A possible explanation for these inconsistent results might be that PolyPhen2 integrates information about protein structure and biochemical properties into their predictions of deleteriousness while SIFT and Panther do not.

Inconsistent predictions were also observed when using the nucleotide-sequence based tools PhastCons and PhyloP. Consistent results were obtained for nine and eight non-synonymous variants in *TMEM132D* and *SLC6A15* respectively. While PhyloP considers each nucleotide independently to estimate the evolutionary conservation score at that position, PhastCons considers also the scores of its neighboured nucleotides. These different approaches might result in different predictions of deleteriousness.

4.4.2 Experimental functional annotation

Experimental functional characterization of genetic variants can only be carried out when the function of the protein is known. This prerequisite sounds trivial, but unfortunately, a large number of human proteins are lacking sufficient functional annotation to design an experimental assay. For instance, the molecular function of TMEM132D is still unknown. It has been suggested that TMEM132D may serve as cell surface marker for the differentiation of oligodendrocytes [328]. Other data have been shown that TMEM132D is predominantly expressed in neurons and co-localized with actin filaments [329]. Due to

the lack of a measurable property that associates with function, experimental assessment of functional consequences of genetic variants was not possible.

In contrast, the SLC6A15 protein is known to transport neutral amino acids into predominantly neuronal cells so that an uptake assay could be performed [7]. The two rare non-synonymous variants T49A and A400V in the *SLC6A15* gene were shown to be associated with a significantly increased ^{3}H proline uptake in HEK cells. High levels of proline were identified to be neurotoxic and have been associated with symptoms of the central nervous system (CNS) such as seizures and mental retardation [330]. Although the uptake measurements were only performed using proline, this amino acid is, of course, not the only substrate for SLC6A15. For instance, the neutral amino acids leucine and methionine are also transported into the cell via the SLC6A15 transporter. Leucine is a major donor of nitrogen for the synthesis of the amino acid glutamate and the neurotransmitter GABA [331]. Therefore, an alteration in leucine uptake could have impact on glutamatergic transmission which is connected to psychiatric disorders [332]. A previously published study, which showed that SLC6A15 is expressed in glutamatergic and GABAergic neurons, supports this hypothesis [333]. Methionine is a precursor of S-adenosylmethionine (SAM) which is the major methyl group donor in humans. SAM transfers methyl groups to different substrates, including DNA nucleotides and histones. SAM metabolism has been associated with different diseases, including psychiatric disorders such as MDD [334].

Alterations in amino acid uptake might be explained by several cellular mechanisms, including altered transporter velocity, gene expression, membrane localization or protein stability. The fact that fluorescence imaging showed no alterations in cellular sub-localization of the protein, and no differences in protein levels between WT and mutant cells, supports the assumption that rare non-synonymous variants in *SLC6A15* lead to functional rather than quantitative changes of the transporter.

Although the proline uptake experiments demonstrated that T49A and A400V significantly increase the level of this amino acid in HEK cells, for T49A only Panther and PhyloP, and for A400V only PhastCons did predict any influence on protein function. This discrepancy is not surprising, as the ENCODE project showed that the correlation between deleterious estimates derived from computational evolutionary annotation, and estimates obtained from experiments is only modest [243,335]. One reason for this modest correlation might be that genetic variants that are biochemically functional do not necessarily have to be biologically relevant so that the phenotype of interest does not have to be affected [243,336]. Indeed, even though variants in *SLC6A15* alter proline

uptake, it can not automatically be concluded that altered protein function is associated with increased or decreased risk for MDD.

4.5 Overall conclusion and outlook

In this thesis, genetic and functional characterization of two candidate genes for common complex psychiatric disorders was performed. For the *TMEM132D* locus, an overrepresentation of putatively functional relevant rare and common variants in controls compared to PD patients could be observed. These results confirm that both rare and common variants within the same gene contribute to disease susceptibility and that rare variants might indeed explain a proportion of the missing heritability. Unfortunately, burden testing is prone to erroneous association, with false positive rates between 50 and 98% [337-339]. Thus, replication in larger independent samples will be required to confirm the robustness of the reported associations of PD with common and rare functional relevant variants in *TMEM132D*.

In addition, an increased number of private variants in *TMEM132D* was present in AD patients compared to controls. This result leads to the suggestion that variants within this gene might be both protective against and risk-increasing for anxiety-related disorders. Unfortunately, case-control studies can only determine associations of genetic variants with a disease, but not which of the associated variants the causal one is. To test the causal relationship between these private variants and AD, this study needs to be supplemented by family studies. The fact that the function of the TMEM132D protein is still unknown further complicates *in vitro* and *in vivo* experimental functional analysis so that, currently, biochemical relevance of these private variants in TMEM132D can not be assessed.

For the *SLC6A15* locus, significant associations with MDD were not observed, neither for common or rare variants alone nor for a combination between both. However, two rare non-synonymous variants were identified to increase the activity of the amino acid transporter without changing the sub-cellular localization of the SLC6A15 protein. These data lead to the suggestion that rare variants in *SLC6A15* might influence the biochemical function of the amino acid transporter. In order to assess the putative relevance of these observed biochemical differences in SLC6A15 transporter activity on neurobiological phenotypes and ultimately MDD, additional experiments in neuronal cells lacking endogenous *SLC6A15*, or humanized transgenic animals are suggested. Additional experimental functional analyses are highly recommended as variants, which

were identified to be biochemically relevant, do not automatically have to be biologically relevant and may not affect the phenotype of interest.

In the future, larger sample sizes are required to detect significant associations of common and rare variants with disease susceptibility for case-control association studies. Although burden testing increases the power to detect genome-wide significant associations, the required sample size is still high, comprising at least several thousands of cases and controls which is only feasible in cooperation with multiple research institutes [340]. Furthermore, studies investigating gene-environment interaction, gene-gene interaction and the epigenome will be necessary to fully understand the mechanisms underlying complex diseases. The risk, which is conveyed by a specific variant, may only be unmasked with exposure to stress or trauma, so that a strict case-control design may be insufficient. This thesis also highlights the need for functional characterization of genetic variants as association alone does not have to mean causality. If possible, experimental validations should be performed to assess possible functionality as computational tools only give insufficient information. In addition, combining variants with regard to their functional relevance and testing this marker set for association with disease has been shown to be the most promising approach for the identification of disease susceptibility variants. For all studies, which will be performed in future, it should always be kept in mind that the phenotype of an individual is the result of a complex interplay between genome, epigenome and environment so that multiple factors are likely to contribute to disease susceptibility.

5. REFERENCES

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106: 9362-9367.
2. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation 118: 1590-1605.
3. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21.
4. Erhardt A, Czibere L, Roeske D, Lucae S, Unschuld PG, et al. (2011) TMEM132D, a new candidate for anxiety phenotypes: evidence from human and mouse studies. Molecular Psychiatry 16: 647-663.
5. Kohli MA, Lucae S, Saemann PG, Schmidt MV, Demirkan A, et al. (2011) The Neuronal Transporter Gene SLC6A15 Confers Risk to Major Depression. Neuron 70: 252-265.
6. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nature Genetics 41: 56-65.
7. Broer A, Tietze N, Kowalczuk S, Chubb S, Munzinger M, et al. (2006) The orphan transporter v7-3 (slc6a15) is a Na+-dependent neutral amino acid transporter (B(0)AT2). Biochemical Journal 393: 421-430.
8. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics 33: 228-237.
9. Pringsheim T, Wiltshire K, Day L, Dykeman J, Steeves T, et al. (2012) The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. Movement Disorders 27: 1083-1091.
10. Salvatore D, Buzzetti R, Baldo E, Furnari ML, Lucidi V, et al. (2012) An overview of international literature from cystic fibrosis registries. Part 4: Update 2011. Journal of Cystic Fibrosis 11: 480-493.
11. Blau N, van Spronsen FJ, Levy HL (2010) Phenylketonuria. Lancet 376: 1417-1427.
12. Lubs HA, Stevenson RE, Schwartz CE (2012) Fragile X and X-Linked Intellectual Disability: Four Decades of Discovery. American Journal of Human Genetics 90: 579-590.
13. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, et al. (2003) Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. Science 301: 386-389.
14. Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448: 591-597.
15. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
16. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, et al. (2011) Genome-wide association study identifies five new schizophrenia loci. Nature Genetics 43: 969-977.
17. Price AL, Marzani-Nissen GR (2012) Bipolar Disorders: A Review. American Family Physician 85: 483-493.
18. Nagel RL (2001) Pleiotropic and epistatic effects in sickle cell anemia. Current Opinion in Hematology 8: 105-110.
19. Merlo CA, Boyle MP (2003) Modifier genes in cystic fibrosis lung disease. Journal of Laboratory and Clinical Medicine 141: 237-241.
20. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250: 1684-1689.

21. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, et al. (1995) Identification of the breast cancer susceptibility gene *BRCA2*. Nature 378: 789-792.
22. Besenbacher S, Mailund T, Schierup MH (2012) Association Mapping and Disease: Evolutionary Perspectives. Methods in Molecular Biology. pp. 275-290.
23. Murken J, Griemm T, Holinski-Feder E (2006) Taschenlehrbuch Humangenetik. Stuttgart: Georg Thieme Verlag.
24. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era - concepts and misconceptions. Nature Reviews Genetics 9: 255-266.
25. Falconer DS (1965) Inheritance of liability to certain diseaes estimated from incidence among relatives. Annals of Human Genetics 29: 51-76.
26. Althoff RR, Faraone SV, Rettew DC, Morley CP, Hudziak JJ (2005) Family, twin, adoption, and molecular genetic studies of juvenile bipolar disorder. Bipolar Disorders 7: 598-609.
27. Shih RA, Belmonte PL, Zandi PP (2004) A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. International Review of Psychiatry 16: 260-283.
28. MacGregor AJ, Snieder H, Schork NJ, Spector TD (2000) Twins - novel uses to study complex traits and genetic diseases. Trends in Genetics 16: 131-134.
29. Collins FS, Lander ES, Rogers J, Waterston RH, Int Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.
30. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.
31. Lander ES, Int Human Genome Sequencing C, Linton LM, Birren B, Nusbaum C, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
32. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. Plos Biology 5: 2113-2144.
33. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nature Reviews Genetics 10: 241-251.
34. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. (2007) Completing the map of human genetic variation. Nature 447: 161-165.
35. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
36. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29: 308-311.
37. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nature Reviews Genetics 7: 85-97.
38. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research 19: 1527-1541.
39. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420-426.
40. Alkan C, Coe BP, Eichler EE (2011) Applications of next-generation sequencing Genome structural variation discovery and genotyping. Nature Reviews Genetics 12: 363-376.
41. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.
42. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nature Genetics 37: 727-732.
43. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

44. Hettema JM, Prescott CA, Myers JM, Neale MC, Kendler KS (2005) The structure of genetic and environmental risk factors for anxiety disorders in men and women. Archives of General Psychiatry 62: 182-189.
45. Sullivan PF, Neale MC, Kendler KS (2000) Genetic epidemiology of major depression: Review and meta-analysis. American Journal of Psychiatry 157: 1552-1562.
46. Wray N, Gottesman I (2012) Using summary data from the Danish National Registers to estimate heritabilities for Schizophrenia, Bipolar Disorder and Major Depressive Disorder. Frontiers in Genetics 3: 118.
47. Faraone SV, Perlis RH, Doyle AE, Smoller JW, Goralnick JJ, et al. (2005) Molecular genetics of attention-deficit/hyperactivity disorder. Biological Psychiatry 57: 1313-1323.
48. Elston RC (1998) Linkage and association. Genetic Epidemiology 15: 565-576.
49. Kullo IJ, Ding K (2007) Mechanisms of Disease: the genetic basis of coronary heart disease. Nature Clinical Practice Cardiovascular Medicine 4: 558-569.
50. Morgan TH (1911) Random segregation versus coupling in Mendelian inheritance. Science 34: 384-384.
51. Morgan TH (1915) Localization of the hereditary material in the germ cells. Proceedings of the National Academy of Sciences of the United States of America 1: 420-429.
52. Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology 14: 43-59.
53. Baron M (2001) The search for complex disease genes: fault by linkage or fault by association? Molecular Psychiatry 6: 143-149.
54. Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. Human Molecular Genetics 9: 2403-2408.
55. Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences of the United States of America 94: 1041-1046.
56. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.
57. Goate A, Chartierharlin MC, Mullan M, Brown J, Crawford F, et al. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349: 704-706.
58. Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, et al. (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. Nature 376: 775-778.
59. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, et al. (1995) Cloning of a gene bearing missense mutations in early onset familial Alzheimer's disease. Nature 375: 754-760.
60. Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22: 139-144.
61. Landegren U, Nilsson M, Kwok PY (1998) Reading bits of genetic information: Methods for single-nucleotide polymorphism analysis. Genome Research 8: 769-776.
62. Knight JC (2003) Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. Clinical Science 104: 493-501.
63. Georges M, Coppieters W, Charlier C (2007) Polymorphic miRNA-mediated gene regulation: contribution to phenotypic variation and disease. Current Opinion in Genetics & Development 17: 166-176.
64. Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, et al. (2010) Single-Nucleotide Polymorphisms Inside MicroRNA Target Sites Influence Tumor Susceptibility. Cancer Research 70: 2789-2798.
65. Law AJ, Kleinman JE, Weinberger DR, Weickert CS (2007) Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. Human Molecular Genetics 16: 129-141.

66. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768-772.
67. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in Transcription Factor Binding Among Humans. Science 328: 232-235.
68. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB (2010) Annotating non-coding regions of the genome. Nature Reviews Genetics 11: 559-571.
69. Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9: 477-485.
70. Lewis CM, Knight J (2012) Introduction to genetic association studies. Cold Spring Harbor protocols 2012: 297-306.
71. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6: 95-108.
72. Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature Genetics 29: 229-232.
73. Goldstein DB (2001) Islands of linkage disequilibrium. Nature Genetics 29: 109-111.
74. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
75. Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics 3: 391-396.
76. Zhu M, Zhao S (2007) Candidate gene identification approach: Progress and challenges. International Journal of Biological Sciences 3: 420-427.
77. DeWan A, Liu M, Hartman S, Zhang SS-M, Liu DTL, et al. (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314: 989-992.
78. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.
79. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics 42: 1118-1125.
80. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics 40: 638-645.
81. Visscher PM (2008) Sizing up human height variation. Nature Genetics 40: 489-490.
82. Hardy J, Singleton A (2009) CURRENT CONCEPTS Genomewide Association Studies and Human Disease. New England Journal of Medicine 360: 1759-1768.
83. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
84. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, et al. (2008) Many sequence variants affecting diversity of adult human height. Nature Genetics 40: 609-615.
85. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nature Genetics 40: 584-591.
86. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40: 575-583.
87. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11: 446-450.
88. Lander ES (1996) The new genomics: Global views of biology. Science 274: 536-539.
89. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends in Genetics 17: 502-510.
90. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? American Journal of Human Genetics 69: 124-137.

91. Fearnhead NS, Winney B, Bodmer WF (2005) Rare variant hypothesis for multifactorial inheritance - Susceptibility to colorectal adenomas as a model. Cell Cycle 4: 521-525.
92. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics 40: 695-701.
93. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19: 212-219.
94. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Human Molecular Genetics 17: 156-165.
95. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
96. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nature Genetics 43: 712-714.
97. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. American Journal of Human Genetics 82: 100-112.
98. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, et al. (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452: 638-642.
99. Charlesworth B (2000) Fisher, Medawar, Hamilton and the evolution of aging. Genetics 156: 927-931.
100. Lord C, Cook EH, Leventhal BL, Amaral DG (2000) Autism spectrum disorders. Neuron 28: 355-363.
101. Power RA, Kyaga S, Uher R, MacCabe JH, Langstrom N, et al. (2013) Fecundity of Patients With Schizophrenia, Autism, Bipolar Disorder, Depression, Anorexia Nervosa, or Substance Abuse vs Their Unaffected Siblings. Jama Psychiatry 70: 22-30.
102. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, et al. (2010) Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. Science 329: 841-845.
103. Bulmer MG (1989) Maintenance of genetic variability by mutation selection balance - a childs guide through the jungle. Genome 31: 761-767.
104. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America 107: 961-968.
105. Raychaudhuri S (2011) Mapping Rare and Common Causal Alleles for Complex Human Diseases. Cell 147: 57-69.
106. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nature Genetics 40: 1199-1203.
107. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genetics 40: 1166-1174.
108. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.
109. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nature Genetics 40: 1107-1112.
110. Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature Genetics 41: 25-34.

111. de Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, et al. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nature Genetics 41: 211-215.
112. Stefansson H, Rujescu D, Cichon S, Pietilainen OPH, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232-236.
113. Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DHR, et al. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455: 237-241.
114. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. New England Journal of Medicine 358: 667-675.
115. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nature Genetics 38: 75-81.
116. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. (2006) Common deletion polymorphisms in the human genome. Nature Genetics 38: 86-92.
117. Hinds DA, Kloek AP, Jen M, Chen XY, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nature Genetics 38: 82-85.
118. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. American Journal of Human Genetics 79: 275-290.
119. Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003-1007.
120. Barber JCK (2005) Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. Journal of Medical Genetics 42: 609-629.
121. Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56: 73-82.
122. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics 5: 618-625.
123. Bateson W (1909) Mendel's principles of heredity. Mendel's principles of heredity.
124. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11: 2463-2468.
125. Phillips PC, Johnson NA (1998) The population genetics of synthetic lethals. Genetics 150: 449-458.
126. Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, et al. (2006) Functional epistasis on a common MHC haplotype associated with multiple sclerosis. Nature 443: 574-577.
127. Phillips PC (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9: 855-867.
128. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. Nature Reviews Genetics 8: 437-449.
129. Turton JC, Bullock J, Medway C, Shi H, Brown K, et al. (2011) Investigating Statistical Epistasis in Complex Disorders. Journal of Alzheimers Disease 25: 635-644.
130. Keck ME, Kern N, Erhardt A, Unschuld PG, Ising M, et al. (2008) Combined Effects of Exonic Polymorphisms in CRHR1 and AVPR1B Genes in a Case/Control Study for Panic Disorder. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 147B: 1196-1204.
131. Witte J (1998) Gene-environment interaction. Encyclopedia of Biostatistics. Chichester: Wiley. pp. 1613-1614.
132. Caspi A, Moffitt TE (2006) Opinion - Gene-environment interactions in psychiatry: joining forces with neuroscience. Nature Reviews Neuroscience 7: 583-590.
133. Plomin R, DeFries J, McClearn G, McGuffin P (2001) Behavioral Genetics. New York: W. H. Freeman.
134. Vesell ES (1991) Genetic and environmental factors causing variation in drug response. Mutation Research 247: 241-257.

135. Hunter DJ (2005) Gene-environment interactions in human diseases. Nature Reviews Genetics 6: 287-298.
136. Caspi A, McClay J, Moffitt TE, Mill J, Martin J, et al. (2002) Role of genotype in the cycle of violence in maltreated children. Science 297: 851-854.
137. Narod SA, Goldgar D, CannonAlbright L, Weber B, Moslehi R, et al. (1995) Risk modifiers in carriers of BRCA1 mutations. International Journal of Cancer 64: 394-398.
138. Smith PG, Day NE (1984) The design of case-control studies - the influence of confounding and interaction effects. International Journal of Epidemiology 13: 356-365.
139. Merikangas KR, He J-P, Brody D, Fisher PW, Bourdon K, et al. (2010) Prevalence and Treatment of Mental Disorders Among US Children in the 2001-2004 NHANES. Pediatrics 125: 75-81.
140. Merikangas KR, He J-p, Burstein M, Swanson SA, Avenevoli S, et al. (2010) Lifetime Prevalence of Mental Disorders in U.S. Adolescents: Results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). Journal of the American Academy of Child and Adolescent Psychiatry 49: 980-989.
141. Kessler RC, Berglund P, Demler O, Jin R, Walters EE (2005) Lifetime prevalence and age-of-onset distributions' of DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62: 593-602.
142. Weiller E, Bisserbe JC, Maier W, Lecrubier Y (1998) Prevalence and recognition of anxiety syndromes in five European primary care settings - A report from the WHO study on Psychological Problems in General Health Care. British Journal of Psychiatry 173: 18-23.
143. Greenberg PE, Sisitsky T, Kessler RC, Finkelstein SN, Berndt ER, et al. (1999) The economic burden of anxiety disorders in the 1990s. Journal of Clinical Psychiatry 60: 427-435.
144. Kessler R, Greenberg P (2002) The economic burden of anxiety and stress disorders. Neuropsychopharmacology: The fifth generation of progress. pp. 981-992.
145. Boden JM, Fergusson DM, Horwood LJ (2007) Anxiety disorders and suicidal behaviours in adolescence and young adulthood: findings from a longitudinal study. Psychological Medicine 37: 431-440.
146. Costello EJ, Egger HL, Angold A (2005) The developmental epidemiology of anxiety disorders: Phenomenology, prevalence, and comorbidity. Child and Adolescent Psychiatric Clinics of North America 14: 631-648.
147. Albert U, Rosso G, Maina G, Bogetto F (2008) Impact of anxiety disorder comorbidity on quality of life in euthymic bipolar disorder patients: differences between bipolar I and II subtypes. Journal of Affective Disorders 105: 297-303.
148. Hasin DS, Goodwin RD, Stinson FS, Grant BF (2005) Epidemiology of major depressive disorder - Results from the National Epidemiologic Survey on Alcoholism and Related Conditions. Archives of General Psychiatry 62: 1097-1106.
149. Amies PL, Gelder MG, Shaw PM (1983) Social phobia - a comparative clinical study. British Journal of Psychiatry 142: 174-179.
150. Kessler RC, Stang PE, Wittchen HU, Ustun TB, Roy-Burne PP, et al. (1998) Lifetime panic-depression comorbidity in the National Comorbidity Survey. Archives of General Psychiatry 55: 801-808.
151. Gregory AM, Caspi A, Moffitt TE, Koenen K, Eley TC, et al. (2007) Juvenile mental health histories of adults with anxiety disorders. American Journal of Psychiatry 164: 301-308.
152. Noyes R, Crowe RR, Harris EL, Hamra BJ, McChesney CM, et al. (1986) Relationship between panic disorder and agoraphobia: a family study. Archives of General Psychiatry 43: 227-232.
153. Mendlewicz J, Papadimitriou G, Wilmotte J (1993) Family study of panic disorder: comparison with generalized anxiety disorder, major depression and normal subjects. Psychiatric Genetics 3: 73-78.

154. Horwath E, Wolk SI, Goldstein RB, Wickramaratne P, Sobin C, et al. (1995) Is the comorbidity between social phobia and panic disorder due to familial cotransmission or other factors? Archives of General Psychiatry 52: 574-582.
155. Fyer AJ, Mannuzza S, Chapman TF, Lipsitz J, Martin LY, et al. (1996) Panic disorder and social phobia: Effects of comorbidity on familial transmission. Anxiety 2: 173-178.
156. Goldstein RB, Wickramaratne PJ, Horwath E, Weissman MM (1997) Familial aggregation and phenomenology of 'early'-onset (at or before age 20 years) panic disorder. Archives of General Psychiatry 54: 271-278.
157. Hettema JM, Neale MC, Kendler KS (2001) A review and meta-analysis of the genetic epidemiology of anxiety disorders. American Journal of Psychiatry 158: 1568-1578.
158. Skre I, Onstad S, Torgersen S, Lygren S, Kringlen E (1993) A twin study of DSM-III-R anxiety disorders. Acta Psychiatrica Scandinavica 88: 85-92.
159. Gelernter J, Bonvicini K, Page G, Woods SW, Goddard AW, et al. (2001) Linkage genome scan for loci predisposing to panic disorder or agoraphobia. American Journal of Medical Genetics 105: 548-557.
160. Fyer AJ, Hamilton SP, Durner M, Haghighi F, Heiman GA, et al. (2006) A third-pass genome scan in panic disorder: Evidence for multiple susceptibility loci. Biological Psychiatry 60: 173-178.
161. Kaabi B, Gelernter J, Woods SW, Goddard A, Page GP, et al. (2006) Genome scan for loci predisposing to anxiety disorders using a novel multivariate approach: Strong evidence for a chromosome 4 risk locus. American Journal of Human Genetics 78: 543-553.
162. Knowles JA, Fyer AJ, Vieland VJ, Weissman MM, Hodge SE, et al. (1998) Results of a genome-wide genetic screen for panic disorder. American Journal of Medical Genetics 81: 139-147.
163. Thorgeirsson TE, Oskarsson H, Desnica N, Kostic JP, Stefansson JG, et al. (2003) Anxiety with panic disorder linked to chromosome 9q in Iceland. American Journal of Human Genetics 72: 1221-1230.
164. Smoller JW, Acierno JS, Rosenbaum JF, Biederman J, Pollack MH, et al. (2001) Targeted genome screen of panic disorder and anxiety disorder proneness using homology to murine QTL regions. American Journal of Medical Genetics 105: 195-206.
165. Weissman MM, Fyer AJ, Haghighi F, Heiman G, Deng ZM, et al. (2000) Potential panic disorder syndrome: Clinical and genetic linkage evidence. American Journal of Medical Genetics 96: 24-35.
166. Middeldorp CM, Hottenga JJ, Slagboom PE, Sullivan PF, de Geus EJC, et al. (2008) Linkage on chromosome 14 in a genome-wide linkage study of a broad anxiety phenotype. Molecular Psychiatry 13: 84-89.
167. Maron E, Hettema JM, Shlik J (2010) Advances in molecular genetics of panic disorder. Molecular Psychiatry 15: 681-701.
168. Duncan LE, Keller MC (2011) A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry. American Journal of Psychiatry 168: 1041-1049.
169. Domschke K, Reif A (2012) Behavioral genetics of affective and anxiety disorders. Current topics in behavioral neurosciences 12: 463-502.
170. McGrath LM, Weill S, Robinson EB, MacRae R, Smoller JW (2012) Bringing a developmental perspective to anxiety genetics. Development and Psychopathology 24: 1179-1193.
171. Chen JS, Lipska BK, Halim N, Ma QD, Matsumoto M, et al. (2004) Functional analysis of genetic variation in catechol-o-methyltransferase (COMT): Effects on mRNA, protein, and enzyme activity in postmortem human brain. American Journal of Human Genetics 75: 807-821.
172. Lachman HM, Morrow B, Shprintzen R, Veit S, Parsia SS, et al. (1996) Association of codon 108/158 catechol-O-methyltransferase gene polymorphism with the

psychiatric manifestations of velo-cardio-facial syndrome. American Journal of Medical Genetics 67: 468-472.
173. Domschke K, Deckert J, O'Donovan MC, Glatt SJ (2007) Meta-analysis of COMT val158met in panic disorder: Ethnic heterogeneity and gender specificity. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 144B: 667-673.
174. Mossner R, Daniel S, Albert D, Heils A, Okladnova O, et al. (2000) Serotonin transporter function is modulated by brain-derived neurotrophic factor (BDNF) but not nerve growth factor (NGF). Neurochemistry International 36: 197-202.
175. Carvalho AL, Caldeira MV, Santos SD, Duarte CB (2008) Role of the brain-derived neurotrophic factor at glutamatergic synapses. British Journal of Pharmacology 153: 310-324.
176. Guillin O, Diaz J, Carroll P, Griffon N, Schwartz JC, et al. (2001) BDNF controls dopamine D-3 receptor expression and triggers behavioural sensitization. Nature 411: 86-89.
177. Hashimoto K (2007) BDNF variant linked to anxiety-related behaviors. Bioessays 29: 116-119.
178. Chen Z-Y, Jing D, Bath KG, Ieraci A, Khan T, et al. (2006) Genetic variant BDNF (Val66Met) polymorphism alters anxiety-related behavior. Science 314: 140-143.
179. Frustaci A, Pozzi G, Gianfagna F, Manzoli L, Boccia S (2008) Meta-Analysis of the Brain-Derived Neurotrophic Factor Gene (BDNF) Val66Met Polymorphism in Anxiety Disorders and Anxiety-Related Personality Traits. Neuropsychobiology 58: 163-170.
180. Heils A, Teufel A, Petri S, Stober G, Riederer P, et al. (1996) Allelic variation of human serotonin transporter gene expression. Journal of Neurochemistry 66: 2621-2624.
181. Greenberg BD, Tolliver TJ, Huang SJ, Li Q, Bengel D, et al. (1999) Genetic variation in the serotonin transporter promoter region affects serotonin uptake in human blood platelets. American Journal of Medical Genetics 88: 83-87.
182. Sen S, Burmeister M, Ghosh D (2004) Meta-analysis of the association between a serotonin transporter promoter polymorphism (5-HTTLPR) and anxiety-related personality traits. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 127B: 85-89.
183. Blaya C, Salum GA, Lima MS, Leistner-Segal S, Manfro GG (2007) Lack of association between the Serotonin Transporter Promoter Polymorphism (5-HTTLPR) and Panic Disorder: a systematic review and meta-analysis. Behavioral and Brain Functions 3.
184. Otowa T, Yoshida E, Sugaya N, Yasuda S, Nishimura Y, et al. (2009) Genome-wide association study of panic disorder in the Japanese population. Journal of Human Genetics 54: 122-126.
185. Otowa T, Tanii H, Sugaya N, Yoshida E, Inoue K, et al. (2010) Replication of a genome-wide association study of panic disorder in a Japanese population. Journal of Human Genetics 55: 91-96.
186. Landgraf R, Kessler MS, Bunck M, Murgatroyd C, Spengler D, et al. (2007) Candidate genes of anxiety-related behavior in HAB/LAB rats and mice: Focus on vasopressin and glyoxalase-I. Neuroscience and Biobehavioral Reviews 31: 89-102.
187. Erhardt A, Akula N, Schumacher J, Czamara D, Karbalai N, et al. (2012) Replication and meta-analysis of TMEM132D gene variants in panic disorder. Translational Psychiatry 2.
188. Mathers CD, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. Plos Medicine 3.
189. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, et al. (2003) The epidemiology of major depressive disorder - Results from the National Comorbidity Survey Replication (NCS-R). Jama-Journal of the American Medical Association 289: 3095-3105.
190. Blair-West GW, Cantor CH, Mellsop GW, Eyeson-Annan ML (1999) Lifetime suicide risk in major depression: sex and age determinants. Journal of Affective Disorders 55: 171-178.

191. Northoff G, Wiebking C, Feinberg T, Panksepp J (2011) The 'resting-state hypothesis' of major depressive disorder-A translational subcortical-cortical framework for a system disorder. Neuroscience and Biobehavioral Reviews 35: 1929-1945.
192. Zisook S, Lesser I, Stewart JW, Wisniewski SR, Balasubramani GK, et al. (2007) Effect of age at onset on the course of major depressive disorder. American Journal of Psychiatry 164: 1539-1546.
193. Kessler RC, McGonagle KA, Zhao SY, Nelson CB, Hughes M, et al. (1994) Lifetime and 12 month prevalence of DSM-III-R psychiatric disorders in the United States - Results from the National Comorbidity Survey. Archives of General Psychiatry 51: 8-19.
194. Keitner GI, Ryan CE, Miller IW, Kohn R, Epstein NB (1991) 12 month outcome of patients with major depression and comorbid psychiatric or medical illness (compound depression). American Journal of Psychiatry 148: 345-350.
195. Fava M, Uebelacker LA, Alpert JE, Nierenberg AA, Pava JA, et al. (1997) Major depressive subtypes and treatment response. Biological Psychiatry 42: 568-576.
196. Weissman MM, Gammon GD, John K, Merikangas KR, Warner V, et al. (1987) Children of depressed parents. Archives of General Psychiatry 44: 847-853.
197. Harrington R, Rutter M, Weissman M, Fudge H, Groothues C, et al. (1997) Psychiatric disorders in the relatives of depressed probands .1. Comparison of prepubertal, adolescent and early adult onset cases. Journal of Affective Disorders 42: 9-22.
198. Nurnberger JI, Foroud T, Flury L, Su J, Meyer ET, et al. (2001) Evidence for a locus on chromosome 1 that influences vulnerability to alcoholism and affective disorder. American Journal of Psychiatry 158: 718-724.
199. Zubenko GS, Hughes HB, Maher BS, Stiffler JS, Zubenko WN, et al. (2002) Genetic linkage of region containing the CREB1 gene to depressive disorders in women from families with recurrent, early-onset, major depression. American Journal of Medical Genetics 114: 980-987.
200. Levinson DF (2006) The genetics of depression: A review. Biological Psychiatry 60: 84-92.
201. Holmans P, Zubenko GS, Crowe RR, DePaulo JR, Scheftner WA, et al. (2004) Genomewide significant linkage to recurrent, early-onset major depressive disorder on chromosome 15q. American Journal of Human Genetics 74: 1154-1167.
202. McGuffin P, Knight J, Breen G, Brewster S, Boyd PR, et al. (2005) Whole genome linkage scan of recurrent depressive disorder from the depression network study. Human Molecular Genetics 14: 3337-3345.
203. Camp NJ, Lowry MR, Richards RL, Plenk AM, Carter C, et al. (2005) Genome-wide linkage analyses of extended Utah pedigrees identifies loci that influence recurrent, early-onset major depression and anxiety disorders. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 135B: 85-93.
204. Breen G, Webb BT, Butler AW, van den Oord EJCG, Tozzi F, et al. (2011) A Genome-Wide Significant Linkage for Severe Depression on Chromosome 3: The Depression Network Study. American Journal of Psychiatry 168: 840-847.
205. Pergadia ML, Glowinski AL, Wray NR, Agrawal A, Saccone SF, et al. (2011) A 3p26-3p25 Genetic Linkage Finding for DSM-IV Major Depression in Heavy Smoking Families. American Journal of Psychiatry 168: 848-852.
206. Lohoff FW (2010) Overview of the Genetics of Major Depressive Disorder. Current Psychiatry Reports 12: 539-546.
207. Savitz J, Lucki I, Drevets WC (2009) 5-HT1A receptor function in major depressive disorder. Progress in Neurobiology 88: 17-31.
208. Alcaro A, Panksepp J, Witczak J, Hayes DJ, Northoff G (2010) Is subcortical-cortical midline activity in depression mediated by glutamate and GABA? A cross-species translational approach. Neuroscience and Biobehavioral Reviews 34: 592-605.
209. Lopez-Leon S, Janssens ACJW, Ladd AMG-Z, Del-Favero J, Claes SJ, et al. (2008) Meta-analyses of genetic studies on major depressive disorder. Molecular Psychiatry 13: 772-785.

210. Risch N, Herrell R, Lehner T, Liang K-Y, Eaves L, et al. (2009) Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression A Meta-analysis. Jama-Journal of the American Medical Association 301: 2462-2471.
211. Sullivan PF, de Geus EJC, Willemsen G, James MR, Smit JH, et al. (2009) Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Molecular Psychiatry 14: 359-375.
212. Lewis CM, Ng MY, Butler AW, Cohen-Woods S, Uher R, et al. (2010) Genome-Wide Association Study of Major Recurrent Depression in the UK Population. American Journal of Psychiatry 167: 949-957.
213. Muglia P, Tozzi F, Galwey NW, Francks C, Upmanyu R, et al. (2010) Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. Molecular Psychiatry 15: 589-601.
214. Rietschel M, Mattheisen M, Frank J, Treutlein J, Degenhardt F, et al. (2010) Genome-Wide Association-, Replication-, and Neuroimaging Study Implicates HOMER1 in the Etiology of Major Depression. Biological Psychiatry 68: 578-585.
215. Shi J, Potash JB, Knowles JA, Weissman MM, Coryell W, et al. (2011) Genome-wide association study of recurrent early-onset major depressive disorder. Molecular Psychiatry 16: 193-201.
216. Wray NR, Pergadia ML, Blackwood DHR, Penninx BWJH, Gordon SD, et al. (2012) Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Molecular Psychiatry 17: 36-48.
217. Holsboer F (2000) The corticosteroid receptor hypothesis of depression. Neuropsychopharmacology 23: 477-501.
218. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74.
219. Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-Scale and automated DNA sequence determination. Science 254: 59-67.
220. Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.
221. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145.
222. Altmann A, Quast C, Weber P (2013) Detecting rare variants for psychiatric disorders using next generation sequencing: a methods primer. Current psychiatry reports 15: 333-333.
223. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42: 30-35.
224. Maxmen A (2011) Exome Sequencing Deciphers Rare Diseases. Cell 144: 635-637.
225. Stratton M (2008) Genome resequencing and genetic variation. Nature Biotechnology 26: 65-66.
226. Metzker ML (2010) Applications of Next-Generation Sequencing technologies - the next generation. Nature Reviews Genetics 11: 31-46.
227. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences of the United States of America 100: 8817-8822.
228. Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T (2006) DNA ligases: Structure, reaction mechanism, and function. Chemical Reviews 106: 687-699.
229. Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. Science 241: 1077-1080.
230. Frank RAW, McRae AF, Pocklington AJ, van de lagemaat LN, Navarro P, et al. (2011) Clustered Coding Variants in the Glutamate Receptor Complexes of Individuals with Schizophrenia and Bipolar Disorder. Plos One 6.

231. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nature Genetics 42: 684-687.
232. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics 43: 1193-1201.
233. Asimit J, Zeggini E (2010) Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics, Vol 44 44: 293-308.
234. Hamilton M (1959) The assessment of anxiety-states by rating. British Journal of Medical Psychology 32: 50-55.
235. Hamilton M (1960) A rating scale for depression. Journal of Neurology Neurosurgery and Psychiatry 23: 56-62.
236. Bandelow B (1999) Panic and Agoraphobia Scale (PAS). Göttingen, Bern, Toronto, Seattle: Hogrefe and Huber Publishers.
237. Wittchen H, H P (1997) DIA-X-Interviews: Manual für Screening-Verfahren und Interview. Frankfurt: Swets & Zeitlinger.
238. Quast C, Altmann A, Weber P, Arloth J, Bader D, et al. (2012) Rare Variants in TMEM132D in a Case-Control Sample for Panic Disorder. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 159B: 896-907.
239. Hennings JM, Owashi T, Binder EB, Horstmann S, Menke A, et al. (2008) Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - Findings from the Munich Antidepressant Response Signature (MARS) project. Journal of Psychiatric Research 43: 215-229.
240. Ising M, Lucae S, Binder EB, Bettecken T, Uhr M, et al. (2009) A Genomewide Association Study Points to Multiple Loci That Predict Antidepressant Drug Treatment Outcome in Depression. Archives of General Psychiatry 66: 966-975.
241. Lucae S, Salyakina D, Barden N, Harvey M, Gagne B, et al. (2006) P2RX7, a gene coding for a purinergic ligand-gated ion channel, is associated with major depressive disorder. Human Molecular Genetics 15: 2438-2445.
242. Quast C, Cuboni S, Bader D, Altmann A, Weber P, et al. (2013) Functional coding variants in *SLC6A15*, a possible risk gene for major depression. Plos One 8.
243. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
244. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20.
245. Yuan H-Y, Chiou J-J, Tseng W-H, Liu C-H, Liu C-K, et al. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Research 34: 635-641.
246. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788-793.
247. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15: 901-913.
248. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4: 1073-1082.
249. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines [et al] Chapter 7: Unit7.20-Unit27.20.
250. Thomas PD, Campbell MJ, Kejariwal A, Mi HY, Karlak B, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Research 13: 2129-2141.

251. Altmann A, Weber P, Quast C, Rex-Haffner M, Binder EB, et al. (2011) vipR: variant identification in pooled DNA using R. Bioinformatics 27: 77-84.
252. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
253. David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: Sensitive yet Practical Short Read Mapping. Bioinformatics 27: 1011-1012.
254. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38.
255. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559-575.
256. Gauderman W, Morrison J (2006) QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies.
257. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics 83: 311-321.
258. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65: 220-228.
259. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004.
260. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genetic Epidemiology 20: 4-16.
261. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. Plos Genetics 2: 2074-2093.
262. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40: 646-649.
263. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nature Biotechnology 29: 24-26.
264. Altmann A, Weber P, Quast C, Rex-Haffner M, Binder EB, et al. (2011) vipR: variant identification in pooled DNA using R. Bioinformatics 27: 77-84.
265. Hartl D, Jones E (1998) Genetics - Principles and Analysis. Sudbury, Massachussets: Jones and Bartlett Publishers.
266. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. American Journal of Human Genetics 80: 727-739.
267. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, et al. (2011) A Genome-wide Comparison of the Functional Properties of Rare and Common Genetic Variants in Humans. American Journal of Human Genetics 88: 458-468.
268. Goldstein JL, Brown MS (1979) LDL receptor locus and the genetics of familial hypercholesterolemia. Annual Review of Genetics 13: 259-289.
269. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, et al. (2007) A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. American Journal of Human Genetics 81: 873-883.
270. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, et al. (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nature Genetics 41: 535-543.
271. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, et al. (2008) Genetic Variation in an Individual Human Exome. Plos Genetics 4.
272. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare Alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869-872.
273. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nature Genetics 39: 513-516.

274. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. Science 324: 387-389.
275. Downes K, Pekalski M, Angus KL, Hardy M, Nutland S, et al. (2010) Reduced Expression of IFIH1 Is Protective for Type 1 Diabetes. Plos One 5.
276. Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, et al. (2006) A spectrum of PCSK9 Alleles contributes to plasma levels of low-density lipoprotein cholesterol. American Journal of Human Genetics 78: 410-422.
277. Visscher P. (2012) XX World Congress of Psychiatric Genetics, Hamburg.
278. Mehta D, Quast C, Fasching PA, Seifert A, Voigt F, et al. (2012) The 5-HTTLPR polymorphism modulates the influence on environmental stressors on peripartum depression symptoms. Journal of Affective Disorders 136: 1192-1197.
279. Brown GW, Craig TKJ, Harris TO (2008) Parental maltreatment and proximal risk factors using the Childhood Experience of Care & Abuse (CECA) instrument: A life-course study of adult chronic depression - 5. Journal of Affective Disorders 110: 222-233.
280. Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, et al. (2013) Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. Nature Neuroscience 16: 33-41.
281. Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, et al. (2013) Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. Nature Neuroscience 16: 33-41.
282. Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, et al. (2004) Epigenetic programming by maternal behavior. Nature Neuroscience 7: 847-854.
283. Champagne FA, Weaver ICG, Diorio J, Dymov S, Szyf M, et al. (2006) Maternal care associated with methylation of the estrogen receptor-alpha 1b promoter and estrogen receptor-alpha expression in the medial preoptic area of female offspring. Endocrinology 147: 2909-2915.
284. Seckl JR, Holmes MC (2007) Mechanisms of Disease: glucocorticoids, their placental metabolism and fetal'programming' of adult pathophysiology. Nature Clinical Practice Endocrinology & Metabolism 3: 479-488.
285. Szyf M, McGowan P, Meaney MJ (2008) The social environment and the epigenome. Environmental and Molecular Mutagenesis 49: 46-60.
286. Neff CD, Abkevich V, Potter J, Riley R, Shattuck D, et al. (2010) Evidence for Epistasis Between SLC6A4 and a Chromosome 4 Gene as Risk Factors in Major Depression. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 153B: 321-322.
287. Pezawas L, Meyer-Lindenberg A, Goldman AL, Verchninski BA, Chen G, et al. (2008) Evidence of biologic epistasis between BDNF and SLC6A4 and implications for depression. Molecular Psychiatry 13: 709-716.
288. Lotrich FE, Pollock BG (2004) Meta-analysis of serotonin transporter polymorphisms and affective disorders. Psychiatric Genetics 14: 121-129.
289. Grabe HJ, Schwahn C, Mahler J, Appel K, Schulz A, et al. (2012) Genetic epistasis between the brain-derived neurotrophic factor Val66Met polymorphism and the 5-HTT promoter polymorphism moderates the susceptibility to depressive disorders after childhood abuse. Progress in Neuro-Psychopharmacology & Biological Psychiatry 36: 264-270.
290. Ressler KJ, Bradley B, Mercer KB, Deveau TC, Smith AK, et al. (2010) Polymorphisms in CRHR1 and the Serotonin Transporter Loci: Gene x Gene x Environment Interactions on Depressive Symptoms. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 153B: 812-824.
291. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature Methods 5: 16-18.

292. Out AA, van Minderhout IJHM, Goeman JJ, Ariyurek Y, Ossowski S, et al. (2009) Deep Sequencing to Reveal New Variants in Pooled DNA Samples. Human Mutation 30: 1703-1712.
293. Lee JS, Choi M, Yan X, Lifton RP, Zhao H (2011) On Optimal Pooling Designs to Identify Rare Variants Through Massive Resequencing. Genetic Epidemiology 35: 139-147.
294. Mitsui J, Fukuda Y, Azuma K, Tozaki H, Ishiura H, et al. (2010) Multiplexed resequencing analysis to identify rare variants in pooled DNA with barcode indexing using next-generation sequencer. Journal of Human Genetics 55: 448-455.
295. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36.
296. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G plus C)-biased genomes. Nature Methods 6: 291-295.
297. Kieleczawa J, Mazaika E (2010) Optimization of protocol for sequencing of difficult templates. Journal of biomolecular techniques : JBT 21: 97-102.
298. Harismendy O, Frazer KA (2009) Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques 46: 229-231.
299. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10.
300. Zhu L, Zhang Y, Zhang W, Yang S, Chen J-Q, et al. (2009) Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics 10.
301. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nature Methods 6: 6-12.
302. Bateman A, Quackenbush J (2009) Bioinformatics for Next-Generation Sequencing. Bioinformatics 25: 429-429.
303. Bao S, Jiang R, Kwan W, Wang B, Ma X, et al. (2011) Evaluation of next-generation sequencing software in mapping and assembly (Retracted article. See vol. 56, pg. 687, 2011). Journal of Human Genetics 56: 406-414.
304. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis 615: 28-56.
305. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.
306. Davis EE, Zhang Q, Liu Q, Diplas BH, Davey LM, et al. (2011) TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. Nature Genetics 43: 189-196.
307. Brunham LR, Singaraja RR, Hayden MR (2006) Variations on a gene: Rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis. Annual Review of Nutrition 26: 105-129.
308. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences of the United States of America 103: 1810-1815.
309. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539-543.
310. Spirin V, Schmidt S, Pertsemlidis A, Cooper RS, Cohen JC, et al. (2007) Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol. American Journal of Human Genetics 81: 1298-1303.
311. Madsen BE, Browning SR (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. Plos Genetics 5.

312. Han F, Pan W (2010) A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Human Heredity 70: 42-54.
313. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes.
314. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare Variants Create Synthetic Genome-Wide Associations. Plos Biology 8.
315. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nature Genetics 36: 388-393.
316. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nature Genetics 36: 512-517.
317. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D (2011) Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genetic Epidemiology 35: 790-799.
318. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11: 459-463.
319. Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, et al. (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. Mutation Research-Reviews in Mutation Research 659: 147-157.
320. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337: 1190-1195.
321. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics 12: 628-640.
322. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Human Molecular Genetics 10: 2319-2328.
323. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends in Genetics 16: 198-200.
324. Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. Journal of Molecular Biology 307: 683-706.
325. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The universal protein resource (UniProt). Nucleic Acids Research 33: 154-159.
326. Bao L, Cui Y (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21: 2185-2190.
327. Dobson RJ, Munroe PB, Caulfield MJ, Saqi MAS (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. Bmc Bioinformatics 7.
328. Nomoto H, Yonezawa T, Itoh K, Ono K, Yamamoto K, et al. (2003) Molecular cloning of a novel transmembrane protein MOLT expressed by mature oligodendrocytes. Journal of Biochemistry 134: 231-238.
329. Walser S, Dedic N, Touma C, Floss T, Wurst W, et al. (2011) TMEM132D - A putative cell adhesion molecule involved in panic disorder. AGNP, Munich.
330. Wyse ATS, Netto CA (2011) Behavioral and neurochemical effects of proline. Metabolic Brain Disease 26: 159-172.
331. Yudkoff M, Daikhin Y, Grunstein L, Nissim I, Stern J, et al. (1996) Astrocyte leucine metabolism: Significance of branched-chain amino acid transamination. Journal of Neurochemistry 66: 378-385.
332. Javitt DC (2004) Glutamate as a therapeutic target in psychiatric disorders. Molecular Psychiatry 9: 984-997.
333. Hagglund MGA, Roshanbin S, Lofqvist E, Hellsten SV, Nilsson VCO, et al. (2013) B(0)AT2 (SLC6A15) Is Localized to Neurons and Astrocytes, and Is Involved in Mediating the Effect of Leucine in the Brain. Plos One 8.

334. Papakostas GI (2009) Evidence for S-Adenosyl-L-Methionine (SAM-e) for the Treatment of Major Depressive Disorder. Journal of Clinical Psychiatry 70: 18-22.
335. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research 17: 760-774.
336. Cooper GM, Brown CD (2008) Qualifying the relationship between sequence conservation and molecular function. Genome Research 18: 201-205.
337. Li G, Ferguson J, Zheng W, Lee JS, Zhang X, et al. (2011) Large-scale risk prediction applied to Genetic Analysis Workshop 17 mini-exome sequence data. BMC proceedings 5 Suppl 9: 46.
338. Hu P, Xu W, Cheng L, Xing X, Paterson AD (2011) Pathway-based joint effects analysis of rare genetic variants using Genetic Analysis Workshop 17 exon sequence data. BMC proceedings 5 Suppl 9: 45.
339. Scholz M, Kirsten H (2011) Comparison of scoring methods for the detection of causal genes with or without rare variants. BMC proceedings 5 Suppl 9: 49.
340. Zoellner S (2012) Sampling strategies for rare variant tests in case-control studies. European Journal of Human Genetics 20: 1085-1091.

6. SUPPLEMENTARY TABLES

Table S1 Oligonucleotides used for amplification of the exons and exon-intron boundaries of the *TMEM132D* locus. Adapted from Quast *et al.* [238].

Primer name	Primer sequence	Amplicon position in genome (hg19)	Amplicon length (bp)
1fwd	AGCACACCCACAGTGCTAACTTTATGT	chr12:130383217-130388188	4972
1rev	GAAAGGAAATACCCCCTGTGGATTAAA		
2fwd	AAAAGCAGCCATAAATCCCATATGAAG	chr12:130183383-130188435	5053
2rev	AGTCCACATAGGGGAAAACTGAGAGTC		
3fwd	AGCATTATACATGCAGCTATGCACCTT	chr12:130013162-130018171	5010
3rev	GAAACATCATCTGAATTCCACATAGCC		
4fwd	TTGCTGGTCTGCAGAATATAGATGTGA	chr12:129820476-129825414	4939
4rev	TTCCTGAAGGTTGGTATAGTCCTGGAT		
5fwd	AAGAATAGAGCAGCAAACAAAGTGGAC	chr12:129691713-129696521	4809
5rev	TTTTGATCCTCCCCCTTTAGAGTAGAA		
6fwd	ATTCTCATCACATCATTACATGGCTTG	chr12:129565819-129570797	4979
6rev	TTAAGAAATGTCCTCACCACAACACAC		
7fwd	TCCTTCAGCTCAGCCAAGAAATACTTA	chr12:129561074-129565875	4802
7rev	CTAGACCCCATCTCAGACAAAACCA		
8fwd	GATTCCATGCATCTTCTTGTTTCGTAG	chr12:129555786-129560718	4933
8rev	TTATATGTGCCCCCACTACACATCTTC		

Table S2 Primer pairs used for LR-PCR amplification of the *SLC6A15* gene. Adapted from Quast *et al.* [242].

Primer name	Primer sequence	Amplicon position in genome (hg19)	Amplicon length (bp)
1fwd	TTTTCTCCCCACCAGCCCCCAATCTGCT	chr12:85302558-85313254	10697
1rev	AACAGCTGAGAAAGCCAGGCCCAAACATCA		
2fwd	AAAACGTCTGCTTCTCCTGCTAGAAACCCCA	chr12:85297924-85308787	10864
2rev	CTCCCACACCAATCCCATGTTGGCCATTTT		
3fwd	GTGATCTGTCAGTTCCAAGAAGGTGTT	chr12:85292757-85303571	10815
3rev	AAAAGAGAGCTTGGTGGCTATCAAAAG		
4fwd	AGCCCAAGAATTCCGCCCTTCATTTCTGGAA	chr12:85287367-85297848	10482
4rev	ACTGCTGCTACCTTCTGGTCAAAGCAAACCA		
5fwd	TGATTTGTGAGAAACAAAAAGCAGGAT	chr12:85284231-85293396	9166
5rev	GATGCTGGATAAGAGGCAAAGAAAAGT		
6fwd	CACAACTTGCAAATCCAATCCCGCCCAGTT	chr12:85279692-85289950	10259
6rev	TCTTCGGTGCAGATGAAGTGCAGTGAGTGAT		
7fwd	TGGGTTCCATGAGGACAGACTGTGGCCTATTA	chr12:85273578-85283585	10008
7rev	ACACTACCCATGTGACCTTTCACAGGCTACCT		
8fwd	AGGCAGCCGCCAGGAGTGACAAAGAAT	chr12:85269233-85271176	1944
8rev	AAACCAAGGGGCAGCCAGCAATTCAGTT		
9fwd	ACATATGCTCGGGGCAGAGCACAAACGTAA	chr12:85257053-85268042	10990
9rev	AGAGGACACGCCATTTGCCATTGTTTGCA		
10fwd	ATTTCCTTATCTGCCAAGTGAAACCAT	chr12:85252553-85263883	11331
10rev	TTCCTATCCAAAAAGTGCATAGCTGAA		
11fwd	AGGCACCACATGGCACGTTTTTGCTGT	chr12:85243329-85253392	10064
11rev	TCTTCTCTCACTCTCTTGCCATGGGGGC		

Table S3 Oligonucleotide primers used for site-directed mutagenesis.

The sequence encoding the substituted amino acid is underlined. The changed nucleotide is bold. Adapted from Quast *et al.* [242].

Mutant	Primer name	Primer sequence
hSLC6A15 T49A	hSLC6A15 T49A-fwd	5'-GGCCAGGAAGAGAAAGAT**G**CAGATGTTGAAGAAGG-3'
	hSLC6A15 T49A-rev	5'-CCTTCTTCAACATCTG**C**ATCTTTCTCTTCCTGGCC-3'
hSLC6A15 K227N	hSLC6A15 K227N-fwd	5'-GGGGGCTTAAACTGGAA**C**ATGACCATCTGCTTG-3'
	hSLC6A15 K227N-rev	5'-CAAGCAGATGGTCAT**G**TTCCAGTTTAAGCCCC-3'
hSLC6A15 A400V	hSLC6A15 A400V-fwd	5'-CAACCTTTCAACTGTTACTG**T**AGAAGATTATCATTTAGTTTATGAC-3'
	hSLC6A15 A400V-rev	5'-GTCATAAACTAAATGATAATCTTCT**A**CAGTAACAGTTGAAAGGTTG-3'
hSLC6A15 L421P	hSLC6A15 L421P-fwd	5'-GAAGAGTTTCCTGCTC**C**TCATCTCAATTCCTGTAAAATTG-3'
	hSLC6A15 L421P-rev	5'-CAATTTTACAGGAATTGAGATGA**G**GAGCAGGAAACTCTTC-3'
hSLC6A15 I500T	hSLC6A15 I500T-fwd	5'-GAGGAAAGAAATTCTTACTGTTA**C**CTGTTGTCTTCTGGC-3'
	hSLC6A15 I500T-rev	5'-GCCAGAAGACAACAG**G**TAACAGTAAGAATTTCTTTCCTC-3'
hSLC6A15 N591D	hSLC6A15 N591D-fwd	5'-GCTAGTGTTGTG**G**ATATGGGATTAAGTCCTCCT-3'
	hSLC6A15 N591D-rev	5'-AGGAGGACTTAATCCCATAT**C**CACAACACTAGC-3'
hSLC6A15 A601T	hSLC6A15 A601T-fwd	5'-CTCCTGGCTATAAC**A**CATGGATTGAAGATAAGG-3'
	hSLC6A15 A601T-rev	5'-CCTTATCTTCAATCCATG**T**GTTATAGCCAGGAG-3'
hSLC6A15 E684D	hSLC6A15 E684D-fwd	5'-GGAAAAATACCGAGCGA**C**ATGCCATCTCCAAATTTTG-3'
	hSLC6A15 E684D-rev	5'-CAAAATTTGGAGATGGCAT**G**TCGCTCGGTATTTTTCC-3'
hSLC6A15 G710R	hSLC6A15 G710R-fwd	5'-GGATACTGCTCCCAAT**A**GACGGTATGGAATAGG-3'
	hSLC6A15 G710R-rev	5'-CCTATTCCATACCGTC**T**ATTGGGAGCAGTATCC-3'

7. LIST OF ABBREVIATIONS

^{3}H	tritium
°C	degree Celsius
µg	microgram
µl	microliter
µM	microMolar
AA	Amino Acid
AD	Anxiety Disorder
ADHD	Attention Deficit Hyperactivity Disorder
ANNOVAR	ANNOtation VARiant
bp	base pair
BWA	Burrows-Wheeler Aligner
CAST	Cohort Allelic Sum Test
CDCV	Common Disease-Common Variant
cDNA	complementary DeoxyriboNucleic Acid
CDRV	Common Disease-Rare Variant
Chip-Seq	Chromatin immunoprecipitation-Sequencing
chr	chromosome
cM	centiMorgan
CMC	Combined Multivariate and Collapsing
CNP	Copy Number Polymorphism
CNS	Central Nervous System
CNV	Copy Number Variant
Con	Control
cpm	counts per minute
ddNTP	dideoxyNucleotide TriPhosphate
DMEM	Dulbecco's Modified Eagle Medium
DNA	DeoxyriboNucleic Acid
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders IV
DZ	DiZygotic
eGFP	enhanced Green Fluorescent Protein
ENCODE	ENCyclopedia Of DNA Elements
ePCR	emulsion Polymerase Chain Reaction
ESE	Exonic Splicing Enhancer

ESP	Exome Sequencing Project
ESS	Exonic Splicing Silencer
FCS	Fetal Calf Serum
GABA	Gamma-AminoButyric Acid
GAD	Generalized Anxiety Disorder
GenRED	Genetics of Recurrent Early-Onset Major Depression
GWAS	Genome-Wide Association Study
HAM-A	HAMilton Anxiety scale
HAM-D	HAMilton Depression scale
HDL	High Density Lipoprotein
HEK	Human Embryonic Kidney
HPA	Hypothalamic-Pituitary-Adrenocortical
IC_{50}	Inhibition Concentration (Inhibition 50%)
ICD	International Statistical Classification of Diseases and Related Health Problems
IGV	Integrative Genomics Viewer
kb	kilobase
LB	Lysogeny Broth
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
LMU	Ludwig-Maximilians-University
LR-PCR	Long Range-Polymerase Chain Reaction
M	Molar
MA	Minor Allele
MAF	Minor Allele Frequency
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization-Time Of Flight
MARS	Munich Antidepressant Response Signature
Mb	Megabase
M-CIDI	Munich version of the Composite International Diagnostic Interview
MDD	Major Depressive Disorder
MDS	MultiDimensional Scaling
miRNA	micro RiboNucleic Acid
mM	millimolar
ml	milliliter
MPIP	Max Planck Institute of Psychiatry

MPS	Massively Parallel Sequencing
mRNA	messenger RiboNucleic Acid
MZ	MonoZygotic
NA	Not Assessed
NAHR	Non-Allelic Homologous Recombination
NaOH	sodium hydroxide
NCS	National Comorbidity Survey
ng	nanogram
NGS	Next-Generation Sequencing
NHGRI	National Human Genome Research Institute
np	not polymorphic
NT	NucleoTide
OR	Odds Ratio
PanIC	Panic International Consortium
PAS	Panic and Agoraphobia Scale
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Panic Disorder
PDL	Poly D-Lysine
PFA	ParaFormAldehyde
pg	picogram
pM	picoMolar
PMA	Presence of Minor Alleles
PTSD	Post-Traumatic Stress Disorder
QC	Quality Control
SAM	S-AdenosylMethionine
SCAN	Schedule for Clinical Assessment in Neuropsychiatry
SD	Standard Deviation
SHRiMP	SHort Read Mapping Package
SIFT	Sorting Intolerant From Tolerant
SKID	Structured Clinical Interview for DSM-IV
SMA	Sum of Minor Alleles
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant

SOLiD	Sequencing by Oligonucleotide Ligation and Detection
STR	Simple Tandem Repeat
TFBS	Transcription Factor Binding Site
UTR	UnTranslated Region
VNTR	Variable Number Tandem Repeat
WFA	Workflow Analysis Run
WHO	World Health Organization
WT	Wild Type
WTCCC	Wellcome Trust Case Control Consortium

8. ACKNOWLEDGEMENTS

Behind every PhD student is a number of people who made it possible. This section is dedicated to all those people who contributed to my thesis.

First of all, I would like to thank Professor Florian Holsboer for giving me the opportunity to prepare my PhD thesis at the Max Planck Institute of Psychiatry.

My special thanks to my supervisor Elisabeth B. Binder for supporting me with her vast knowledge and expertise in psychiatric genetics, inspiring me with her fascination for realizing scientific goals and providing me valuable inputs and comments on my work.

Many thanks to André Altmann, Daniel Bader, Simone Röh and Janine Arloth for supporting me in statistics and data processing. Especially the analysis of next-generation sequencing data would not have been possible without the help of my statisticians.

Thanks to Serena Cuboni and Felix Hausch for their contribution to experimental functional annotation of non-synonymous variants in the *SLC6A15* gene.

I would also like to thank Maik Ködel, Susann Sauer and Monika Rex-Haffner for their excellent technical support.

Furthermore, thanks to the clinicians, psychologists and researchers who were responsible for sample recruitment: Susanne Lucae, Angelika Erhardt, Tanja Brückl, Marcus Ising, Hildegard Pfister, Angela Heck and Anna Kopczak.

My gratitude extends to my work colleagues Peter Weber for his help and discussions with regard to next-generation sequencing, Divya Mehta who has always taken the time to read my English manuscripts and especially Anne Löschner who was always open for my questions and problems.

I would like to mention my gratitude towards all my friends and family especially my sister Bianca Quast and my best friends Maria Hain and Stefanie Wehner for always supporting and believing in me through the almost four years of my PhD.

All this would not have been possible without the love and support of my mother and my father who guided me through every step in my life. I am forever indebted to my parents for their encouragement in all stages of my previous career and their financial support which gave me the opportunity to become a biologist.

At last, but not least I owe my deepest gratitude to Stefan Darchinger for always helping me, believing in me and most importantly for his endless patience when it was needed.

9. APPENDIX

9.1 Curriculum vitae

Personal details

Name	Carina Quast
Date of birth	January 24th, 1985
Nationality	German
Email	carina_quast@yahoo.de

Education

01/2010 – 11/2013 **PhD student**
Max Planck Institute of Psychiatry, Munich
Research group „Molecular Genetics of Affective Disorder"
Title of the PhD thesis:
„Genetic and functional characterization of candidate genes for complex psychiatric diseases using next-generation sequencing and cellular uptake assays"

10/2004 - 9/2009 **Degree in Biology** (Grade 1,0)
Qualification – Diploma in Biology (Dipl. Biol)
University of Würzburg

- Primary subject - Biotechnology
- Secondary subjects - Human Genetics, Pharmaceutical Biology

Diploma research thesis (Grade 1,0)
Max Planck Institute of Psychiatry, Munich
Title of the diploma research thesis:
„Validierung von Kandidatengenen aus genomweiten Assoziationsstudien bei Angststörungen"

7/2004 **University-Entrance Diploma** (Grade 1,3)
Wirtschaftsgymnasium, Bad Mergentheim

9.2 List of publications

2013 **Quast C**, Reif A, Brückl, T, Pfister H, Weber H, Mattheisen M, Cichon S, Lang T, Hamm A, Fehm L, Ströhle A, Arolt V, Domschke K, Kircher T, Wittchen U, Pauli P, Gerlach A, Alpers GW, Deckert J, Rupprecht R, Binder EB, Erhardt A.
Gender-specific association of variants in the AKR1C1 gene with dimensional anxiety in patients with panic disorder. Depression and Anxiety: *In Press.*

2013 **Quast C,** Cuboni S, Bader D, Altmann A, Weber P, Arloth J, Röh S, Brückl T, Ising M, Kopczak A, Erhardt A, Hausch F, Lucae S, Binder EB.
Functional coding variants in SLC6A15, a possible risk gene for major depression. Plos One 8(7): e68645.

2013 Altmann A, **Quast C,** Weber P.
Detecting rare variants for psychiatric disorders using next-generation sequencing: a methods primer. Current psychiatry reports 15(1): 333.

2012 **Quast C,** Altmann A, Weber P, Arloth J, Bader D, Heck A, Pfister H, Müller-Myhsok B, Erhardt A, Binder EB.
Rare Variants in TMEM132D in a Case-Control Sample for Panic Disorder. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 159B(8): 896-907.

2012 Mehta D*, **Quast C***, Fasching PA, Seifert A, Voigt F, Beckmann MW, Faschingbauer F, Burger P, Ekici AB, Kornhuber J, Binder EB, Goecke TW.
The 5-HTTLPR polymorphism modulates the influence on environmental stressors on peripartum depression symptoms. Journal of Affective disorders 136: 1192-1197.

2011 Altmann A, Weber P, **Quast C,** Rex-Haffner M, Binder EB, Müller-Myhsok B.
VipR: Variant identification in pooled DNA using R. Bioinformatics 27: 77-84.

2010 **Quast C,** Binder EB, Erhardt A, Czibere L.
Auffinden molekularer Targets für die Therapie von Angststörungen. Laborwelt 4/2010: 21-23.

9.3 Conferences and Workshops

10/2012	„XX World Congress of Psychiatric Genetics" of the International Society of Psychiatric Genetics (ISPG) in Hamburg, Germany
06/2012	„European Human Genetics Conference 2012" of the European Society of Human Genetics (ESHG) in Nürnberg, Germany
10/2011	„27. Symposium der Arbeitsgemeinschaft für Neuropsychopharmakologie und Pharmakopsychiatrie (AGNP)" in Munich, Germany
03/2011	„Workshop on Neuropsychopharmacology for Young Scientists in Europe" of the European College of Neuropsychopharmacology (ECNP) in Nice, France
10/2010	„XVIII World Congress of Psychiatric Genetics" of the International Society of Psychiatric Genetics (ISPG), Athens, Greece

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist.
Die vorgelegte Dissertation wurde weder ganz, noch in wesentlichen Teilen bei einer anderen Prüfungskommission vorgelegt.
Ich habe zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den 10. Juli 2013 *Carina Quast*