Genetics of schizophrenia: Contribution of rare and low-frequency sequence variants in selected genomic regions to disease susceptibility

Dissertation

zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn
vorgelegt von

Fitnat Buket Basmanav

aus

Ankara

Bonn 2014

Anfertigung mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Sven Cichon

2. Gutachter: Prof. Dr. Michael Hoch

Tag der mündlichen Prüfung: 09. Dezember 2014

Erscheinungsjahr: 2015

Dedicated to my parents Secil and Erman Basmanav

TABLE OF CONTENTS

ABBREVIATIONS

1. INTRODUCTION **1**

1.1 Schizophrenia 1

1.2 Models proposed for the genetic architecture of schizophrenia 2

1.2.1 Common disease-common variant model 3

1.2.2 Common disease-rare variant model 3

1.3 Genetic studies of schizophrenia 4

1.3.1 Early approaches and discoveries 4

1.3.2 Genome-wide association studies 6

1.3.3 Copy number variation studies 9

1.3.4 Sequencing studies in schizophrenia 12

1.3.4.1 Region or gene targeted resequencing studies in schizophrenia 12

1.3.4.2 Exome sequencing studies in schizophrenia 13

1.4 The emerging genetic architecture of schizophrenia 15

1.5 Aims of the doctoral thesis 16

2. MATERIALS AND METHODS **18**

2.1 Devices 18

2.2 Chemicals, Reagents, Buffers and Solutions 19

2.3 Commercial systems 20

2.4 Softwares and databases 20

2.5 Study Samples 21

2.6 Methods 24

2.6.1 Isolation and management of nucleic acids 24

2.6.2 Reverse transcription 26

2.6.3 Real-Time Quantitative PCR 27

2.6.4 Sanger sequencing 29

2.6.5 Genotyping based on MassExtend Reaction (Sequenom®) 32

2.6.6 Genome-wide gene expression analysis 36

2.6.7 Variant detection and bioinformatical analysis 39

2.6.8 Statistical analysis 40

3. RESULTS **43**

3.1 Resequencing of the *NRXN1* gene 43

3.2 Resequencing of the 1q21.1 microdeletion region 45

3.2.1 Microarray gene expression analysis of genes at the 1q21.1 microdeletion region in human hippocampus 51
3.2.2 Association analysis of the resequencing data from 1q21.1 microdeletion region 52
3.2.3 Follow-up analyses of *CHD1L* gene 54
3.2.3.1 Genotyping and association analysis of the low-frequency *CHD1L* variants........54
3.2.3.1.1 Functional assessment of rs36008075 55
3.2.3.2 Extended resequencing of *CHD1L* gene and follow-up of individual variants........57
3.2.4 Follow-up analyses of *FMO5* gene 60
3.3 Resequencing of the *TCF4* gene 61
3.3.1 Association analysis of low frequency variants in *TCF4* gene........62
4. DISCUSSION........65
4.1 Rare genetic variation in the *NRXN1* gene and schizophrenia 65
4.2 Rare genetic variation in the 1q21.1 microdeletion region and schizophrenia 70
4.3 Rare genetic variation in the *TCF4* gene and schizophrenia 79
5. SUMMARY........85
6. OUTLOOK........88
7. REFERENCES........90
8. LIST OF PUBLICATIONS........113
9.ATTACHMENTS........115
ACKNOWLEDGMENTS

ABBREVIATIONS

A	Alanine
AA	Amino acid
ABCA13	ATP-binding cassette sub-family A member 13
ABI	Applied Biosystems
ACP6	Lysophosphatidic acid phosphatase type 6
AD	Transcription activation domain
AMBRA1	Activating molecule in BECN1-regulated autophagy protein 1
ANK3	Ankyrin-3
ASD	Autism spectrum disorders
BCL9	B-cell CLL/lymphoma 9 protein
bHLH	Basic helix-loop-helix domain
bp	Base pairs
BRCA1	Breast cancer type 1 susceptibility protein
BRCA2	Breast cancer type 2 susceptibility protein
BRP44	Brain protein 44
C	Cysteine
°C	Degree Celcius
C-	Carboxy-
C10orf26	WW domain binding protein 1-like
Ca2+	Calcium
CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A
CaM	Calmodulin
CaMKII	Calcium/calmodulin-dependent protein kinase II
CCDC68	Coiled-coiled domain containing 68
CDCV	Common disease-common variant
cDNA	Complementary deoxyribonucleic acid
CDRV	Common disease-rare variant
CEU	Utah residents with ancestry from northern and western Europe
CHD1L	Chromodomain-helicase-DNA-binding protein 1-like
Chr	Chromosome
CHRM4	Muscarinic acetylcholine receptor M4
CMC	Combined Multivariate and Collapsing
CNNM2	Metal transporter CNNM2
CNS	Central nervous system
CNV	Copy number variation
CO_2	Carbon dioxide
conc.	Concentration
cRNA	Copy RNA
CSMD1	CUB and sushi domain-containing protein 1
C_t	Cycle threshold
Cy	Cyanine
Cyc	Cyclophilin
CZP1	Zonular pulverulent cataract
D	Aspartate
DBM	Distance-based measure
ddATP	Dideoxyadenosine triphosphate
ddCTP	Dideoxycytidine triphosphate
ddGTP	Dideoxyguanosine triphosphate

ddTTP	Dideoxythymidine triphosphate
ddNTP	Dideoxynucleotide triphosphate
DGKZ	Diacylglycerol kinase zeta
dH_2O	Distilled water
DISC1	Disrupted in schizophrenia 1 protein
DISC2	Homo sapiens disrupted in schizophrenia 2
DLG1	Disks large homolog 1
DLPFC	Dorsolateral prefrontal cortex
DNA	Deoxyribonucleic acid
DMSO	Dimethyl sulfoxide
dNTP	Deoxynucleotide triphosphate
DS	Deletion syndrome
DSM	Diagnostic and Statistical Manual of Mental Disorders
dT	Deoxythymidine
DTT	Dithiothreitol
DPYD	Dihydropyrimidine dehydrogenase [NADP(+)]
E	Glutamate
EBV	Eppstein-Barr virus
e.g.	For example
EPHB1	Ephrin type-B receptor 1
et al.	Et alii
EtBr	Ethidium bromide
EtOH	Ethanol
F	Forward
F	Phenylalanine
Fig.	Figure
FIN	Finnish from Finland
FMO5	dimethylaniline monooxygenase [N-oxide-forming] 5
fMRI	Functional magnetic resonance imaging
FRET	Fluorescence resonance energy transfer
G	Glycine
GABA	Gamma-aminobutyric acid
GBR	British from England and Scotland
GPHN	Gephyrin
GJA5	Gap junction alpha-5 protein
GJA8	Gap junction alpha-8 protein
GPR89B	G protein-coupled receptor 89B
GPR89C	G protein-coupled receptor 89C
GRIK4	Glutamate receptor ionotropic, kainate 4
GWAS/GWASs	Genome-wide association study / Genome-wide association studies
H	Histidine
H_2O	Water
HSF	Human Splicing Finder
I	Isoleucine
ID	Identity number
i.e.	That is
IBS	Iberian populations in Spain
ICD10	International Classification of Diseases
ISC	International schizophrenia consortium

ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3
ITIH4	Inter-alpha-trypsin inhibitor heavy chain H
K	Lysine
KALRN	Kalirin
kb	Kilobase pairs
L	Leucine
LAMA2	Laminin subunit alpha-2
LD	Linkage disequilibrium
LRP1B	Low-density lipoprotein receptor-related protein 1B
LSM1	U6 snRNA-associated Sm-like protein LSm1
M	Methionine
M	Molar
MAF	Minor allele frequency
MALDI-ToF-MS	Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
Mb	Megabase pairs
MDK	Midkine
MEF-2A	Myocyte-specific enhancer factor 2A
mg	Milligram
$MgCl_2$	Magnesium chloride
MHC	Major Histocompatibility Complex
MGS	Molecular Genetics of Schizophrenia Consortium
min	Minute
miR	MicroRNA
mM	Millimolar
MMP16	Matrix metalloproteinase-16
M-PVA	Magnetic poly-vinyl alcohol
mRNA	Messenger ribonucleic acid
µl	Microliter
µM	Micromolar
N	Asparagine
n	Number
N-	Amino-
n.a.	Not available
NBPF11	Neuroblastoma breakpoint family member 11
NDE1	Nuclear distribution protein nudE homolog 1
NDEL1	Nuclear distribution protein nudE-like 1
NLS	Nuclear localization signal
nm	Nanometer
ng	Nanogram
NMD	Non-sense mediated mRNA decay
NMDA	N-methyl-D-aspartate
NPAS3	Neuronal PAS domain-containing protein 3
NRGN	Neurogranin
NT5C2	Cytosolic purine 5'-nucleotidase
NRXN1	Neurexin-1
OD	Optical density
OR	Odds ratio

P	Proline
PAK2	Serine/threonine-protein kinase PAK 2
PCM1	Pericentriolar material 1 protein
PCGEM1	Prostate-specific transcript 1
PCR	Polymerase chain reaction
PDE4B	cAMP-specific 3',5'-cyclic phosphodiesterase 4B
PGC	Schizophrenia Psychiatric Genome-Wide Association Study Consortium
pmol	Picomol
PRKAB2	5'-AMP-activated protein kinase subunit beta-2
PTC	Premature termination codon
PTHS	Pitt Hopkins syndrome
Q	Glutamine
R	Arginine
R	Reverse
RIN	RNA integrity number
RNA	Ribonucleic acid
rpm	Rounds per minute
rRNA	Ribosomal RNA
RT	Room temperature
RT-qPCR	Real-Time quantitative PCR
S	Serine
SAP	Shrimp alkaline phosphatase
SBE	Single-base extension
sec	Second
SKAT	Sequence kernel association test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRI	Solid Phase Reversible Immobilization
T	Threonine
TBE	Tris borate EDTA buffer
TCF4	Transcription factor 4
TRRAP	Transformation/transcription domain-associated protein
TSI	Toscani in Italia
TSPAN18	Tetraspanin-18
U	Unit
UEP	Unextended primer
UTP	Uridine triphosphate
V	Valine
VAV	Guanine nucleotide exchange factor VAV3
VGLUT1	Vesicular glutamate transporter 1
VIPR2	Vasoactive intestinal polypeptide receptor 2
VPS39	Vam6/Vps39-like protein
VRK2	Serine/threonine-protein kinase VRK2
vs.	Versus

W	Tryptophan
WHSC1L1	Histone-lysine N-methyltransferase NSD3
WSS	Weighted-sum statistic
ZNF804A	Zinc finger protein 804A

1. INTRODUCTION

1.1 Schizophrenia

Schizophrenia is a severe chronic neuropsychiatric disorder with a lifetime prevalence and lifetime morbid risk of 0.4% and 0.72%, respectively[1]. The manifestation of the disorder is typically characterized by an admixture of persistent positive and negative symptoms as well as cognitive and motor deficits which result in an impairment of emotional and social behaviour. Some of the symptoms include hallucinations and delusions, distorted perception, disorganized thinking and behaviour, apathy, excessive or slowed motor activity and deficits in different domains of cognition including episodic memory, verbal fluency, processing speed, attention and working memory[2]. The occurrence and severity of different symptoms changes across patients as well as throughout the course of their illness. In the majority of the cases, the onset of psychotic symptoms occurs during adolescence or in early adulthood with a well established earlier age of onset for males in comparison to females[3]. The sustained recovery is only experienced by a minority of the individuals diagnosed with schizophrenia; <14% within the first 5 years of the illness[4] and another 16% later in the illness[5]. Schizophrenia is an incurable however a treatable illness but the current treatments suffer from limited efficacy with the majority of the patients experiencing multiple relapses[4, 6]. The treatment outcomes are extremely variable from case to case due to the high clinical heterogeneity of the illness that are influenced by inter-personal differences including age, gender, race, genetic constitution[7]. In addition, the antipsychotic drugs used in treatment of schizophrenia are known to cause several neurological, metabolic, cardiovascular, gastrointestinal, hematological, genito-urinary and musculoskeletal side effects[7]. Due to the unavailability of consistent and reliable biomarkers, the diagnosis of schizophrenia relies on phenomenology; mainly the symptoms, signs, and course of illness as guided by *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (DSM-IV) or *International Classification of Diseases*, tenth edition (ICD-10)[8].

The pathophysiology of schizophrenia is still not understood but there are widespread yet variable structural, functional and neurochemical alterations reported in the brains of schizophrenia patients. Reduction in grey matter volume at certain brain regions (e.g. cingulate cortex, insula, thalamus, frontal lobe gyri)[9], enlargement of lateral and third ventricular spaces[9], disrupted anatomical connectivity associated to white matter abnormalities[10], inter-regional functional dysconnectivity[11, 12], abnormal activation patterns in several brain regions during performance of executive tasks[13-15] and abnormalities in several neurotransmitter systems (i.e. glutamatergic[16-18], dopaminergic[19, 20], GABAergic[21, 22]) can be listed among the repeatedly reported alterations in patients with schizophrenia. The existence of such widespread alterations can imply disparate pathophysiological models which can account for the well

known clinical heterogeneity in schizophrenia[23]. However it should also be acknowledged that these alterations are not always necessarily distinct entities and some can be linked and converge on a single pathophysiological mechanism such as an imbalance between excitatory and inhibitory neural systems due to glutamatergic and GABAergic dysfunction[3].

Schizophrenia is a multifactorial disorder (also commonly referred to as a “complex disorder”) with joint effects of genetic and environmental risk factors (e.g. urbanicity, cannabis use, developmental trauma, minority group position[24]) playing a role in its aetiology. The genetic component of the disease was demonstrated through family, twin and adoption studies which produced heritability estimates in the range of 64-81%[25, 26]. The disease recurrence risk in a family is known to exponentially increase with increased degree of relatedness to the affected individual where the risk reaches 50% for the monozygotic co-twin of a schizophrenia patient[26]. Schizophrenia does not have a Mendelian form with none of the schizophrenia associated mutations causing the disease per se[27, 28]. In fact there is a highly complex genetic architecture underlying the disease. All the schizophrenia associated variants identified up to date are falling into an effect size spectrum from small to large effect sizes[27, 29] and even for the largest known genetic risk factor (i.e. a microdeletion in chromosomal region 22q11) the rate at which carriers develop the disease is 25–30%[30]. Among the robustly defined schizophrenia associated genetic variants are the several common single nucleotide polymorphisms (SNPs) with small effect sizes and rare large structural variants, namely copy number variants (CNVs) with intermediate to large effect sizes[27, 29]. The efforts to underpin the complete spectrum of genetic risk factors for schizophrenia are vigorously ongoing. The value of revealing the complete spectrum of disease associated loci and alleles lies in the better understanding of the disease pathophysiology. This is consequently expected to lead to development of more effective therapeutics targeting directly the pathophysiological mechanisms. The subsequent goals that are desired to be achieved are still distal but not completely unrealistic and include personally tailored medicine and the possibility of early intervention and prevention by predicting individuals at increased risk of disease[29].

1.2 Models proposed for the genetic architecture of schizophrenia

The concept of ‘genetic architecture of a disease’ compromises 4 main components[31]:

1. The number of risk alleles contributing to the disease in the population
2. The individual frequencies of risk alleles in the population
3. The individual effect sizes of the risk alleles
4. The way the risk alleles act together (independently (additive) or inter-dependently (interacting))

The genetic architecture of schizophrenia has been a major issue of debate initially dominated by two models[32, 33] based on different hypotheses which are not mutually exclusive[34]. One of these is the common disease-common variant (CDCV)[32] hypothesis and the other is the common disease-rare variant (CDRV) [33] hypothesis.

1.2.1 Common disease-common variant model

The original common disease-common variant (CDCV) hypothesis[35, 36] which is today cited as the foundation of the CDCV model suggested that if a disease is common it would be likely that there is a disease causing allele which is much more common than all the other disease causing alleles at the same locus. The initial hypothesis was therefore not accounting for the total number of disease loci and risk variants, the allele frequency spectrum of risk variants in the genome or the effect sizes[37]. Today the CDCV hypothesis is used to define a genetic architecture where the disease is caused by inheritance of multiple relatively common variants each with small individual effect sizes and acting in concert to increase the disease risk[29, 38]. The CDCV hypothesis also fits into a polygenic disease model where a large number of disease relevant genetic loci and risk alleles are expected to be present in the genome[39]. Under this model the disease occurs when an individual harbours a substantial amount of risk alleles, in other words, when a certain threshold of genetic liability is passed[31, 40]. Incidence rates of schizophrenia remain to be persistent despite the deleterious nature of the disease to reproductive fitness[41]. Regarding this paradox, the polygenic model puts forward the argument of reduced visibility of each risk allele to negative selection due to their small individual effect sizes[39].

The first suggestion that a polygenic model can underlie the heritability of schizophrenia came from Gottesman and Shields in 1967[40]. Today the latest evidence from assessment of genome-wide SNP data supports the presence of a polygenic component in the genetic architecture of schizophrenia[27, 39, 42]. A substantial proportion of this polygenic architecture was shown to be attributable to common variants with small individual effect sizes as CDCV predicts, however the contribution of rare variants to schizophrenia is also acknowledged as they are also expected and shown to have a part in the allele frequency/effect size spectrum of a polygenic model[39, 42].

1.2.2 Common disease-rare variant model

The CDCV model is challenged by the common disease-rare variant model (CDRV) which argues that recently arisen, rare, multiple variations are more likely to contribute to the common human disease[43]. In this framework, the disease is caused by different and very rare risk variants (might be even specific to a single case or a family) with large effect sizes each. This model is also referred to as the genetic heterogeneity model since it assumes that each case or family is explained by its own mutation and this must thus account for a huge number of risk

alleles and risk loci spread throughout the entire genome[33]. Genetic heterogeneity has been shown in some other common diseases such as breast cancer where more than thousand individually rare pathogenic mutations in the genes *BRCA1* and *BRCA2* were identified[44]. In the CDRV model the paradigm of high incidence rates of schizophrenia despite reduced fecundity is explained by the suggested continuous occurrences of *de novo* mutations (alterations in genome which are not present in the somatic cells of parents but only in the offspring as a result of a mutation either in a germ cell or in the fertilized egg itself) which are adding sporadic cases of schizophrenia to the population[38]. Supporters of this model in schizophrenia emphasize the association of increased paternal age with schizophrenia with relevance to increased *de novo* germline mutation rates in advanced paternal age[33]. Today it is known that rare structural variants (CNVs) with intermediate to large effect sizes are associated with schizophrenia and they cumulatively are suggested to be present in 1.25%[45] to 2-3%[46] of schizophrenia cases. However these rare variants are also carried by healthy individuals so the development of disease cannot be solely attributed to the carrier status.

Based on the recent genetic findings in schizophrenia, it has become clearer that not one model alone describes the genetic architecture perfectly and that most individuals are rather to have a spectrum of common and rare susceptibility variants.

1.3 Genetic studies of schizophrenia

1.3.1 Early approaches and discoveries

Early studies of schizophrenia genetics were restricted to the technological possibilities of their time and relied mainly on linkage studies, cytogenetic studies and candidate gene association studies.

Linkage studies investigate the co-segregation of genetic markers with disease status in families where there are multiple affected individuals[47]. In other words, it assesses the identical chromosomal regions that are repeatedly shared between the affected individuals in a pedigree with the assumption that these regions would harbour the true disease relevant loci and variants that explain familial transmittance of a disease. Genome-wide linkage scans have been very successful in defining genetic loci underlying monogenic disorders[48]; however when applied to complex disorders like schizophrenia they yielded most of the time not replicable findings due to the underlying genetic heterogeneity and contribution of environmental factors to the occurrence of the disease. Up to date over 30 genome-wide linkage scans and a meta-analysis of 32 independent genome-wide linkage scans of schizophrenia have been performed pointing to different large chromosomal regions spanning hundreds of genes[49].

Cytogenetic studies enable visualization of chromosomes in a eukaryotic cell and detection of microscopic alterations in chromosomes such as aneuploidies (abnormal number of chromosomes) or structural alterations like deletions, duplications, translocations and inversions. There have been several reports where chromosomal abnormalities were detected in individuals with schizophrenia; while some of these reports were single case studies with no direct evidence for a true causality, some were able to perform linkage analysis and show the co-segregation of the abnormality with psychiatric illness within the individual's family[50]. In 1990, St. Claire and colleagues identified a large Scottish pedigree of four generations where a balanced reciprocal translocation between chromosome 1 and chromosome 11 co-segregated with a range of psychiatric disorders (schizophrenia, bipolar disorder and recurrent major depression) in the pedigree. The translocation directly affected two brain expressed genes, *DISC1* and *DISC2* (disrupted in schizophrenia 1 and disrupted in schizophrenia 2 genes; the latter encoding for a non-coding RNA molecule which is antisense to *DISC1*) located at the breakpoint of chromosome 1[51]. After this initial discovery many studies elucidating the biological function of DISC1 have shown that DISC1 indeed acts as a key protein during and after brain development; particularly in the cerebral cortex; by playing a role in multiple cellular processes[52-54]. Also it is the binding partner of many other proteins of importance to brain function and development, some of which have gained support as candidate risk genes (*NDEL1*, *PCM1*, *PDE4B, NDE1*) for psychiatric disorders from independent genetic studies[55]. Since the discovery of the Scottish pedigree, common and rare genetic variation at the locus has extensively been studied in schizophrenia, however apart from the original translocation event no other variant at this locus has gained consistent evidence for contribution to disease susceptibility[29, 56]. Cytogenetic studies in schizophrenia suggested other disease-predisposing mutations in *NPAS3*[57], *GRIK4*[58] and *ABCA13*[59] genes.

Another important discovery in early studies of schizophrenia genetics was the association of schizophrenia with rare recurrent microdeletions at chromosomal region 22q11.2. The microdeletions (either 3 Mb or 1.5 Mb in size) in this region were known to cause the 22q11.2 deletion syndrome (22q11.2DS, also known as velocardiofacial syndrome or DiGeorge syndrome) which manifests itself with a wide range of phenotypes including craniofacial and cardiovascular anomalies, immunodeficiency, short stature, cognitive impairments and behavioural disturbances[60]. Importantly, up to one-third of deletion carriers develop schizophrenia or schizoaffective disorder at late adolescence or early adulthood[61, 62]. Based on these implications; in 1995[63] a sample of 100 schizophrenia patients were screened for interstitial deletions of chromosome 22q11 and 2 of them (2%) were identified to be carriers. This finding was repeatedly confirmed by following studies and it is now well established that rare, recurrent, *de novo* 22q11.2 microdeletions are the strongest genetic risk factor for

schizophrenia known to date and they account for up to 1-2% of all schizophrenia cases and have a major role in continuously introducing new sporadic cases to the population[64, 65]. It is also interesting that there are no major clinical differences between 22q11.2 microdeletion carrier and non-carrier schizophrenia patients with regards to their core schizophrenia phenotype[66] and many 22q11.2 deletion carrier schizophrenia patients are reported to be indistinguishable from non-carrier schizophrenia patients since they have very subtle congenital anomalies and no intellectual disabilities[60]. The shorter and the longer forms of the deletion span, respectively, about 35 and 60 genes, most of which are expressed in brain; however it is not yet known precisely how many and which of these genes are the underlying risk genes for developing psychosis and whether they act in an independent or an interacting fashion.

Genetic association studies compare frequencies of an allele of interest between affected and non-affected individuals where a significantly different allele frequency is suggestive of an etiological relevance of the particular SNP or its location to the disease. Early association studies in schizophrenia were performed for candidate genes whose candidature were based on i) localization in linkage or cytogenetic abnormality regions, ii) biological roles in central nervous system (CNS) and iii) psychopharmacological hypotheses[29]. Retrospectively, most of these studies had very limited power to detect risk variants of small effect size[29]. This was not recognized at that time because the effects of major risk genes were assumed to be larger than they actually appear to be. Consequently, these studies have yielded inconsistent findings not fulfilling the modern criteria for replication[28, 67].

1.3.2 Genome-wide association studies

Genome-wide association studies (GWASs) refer to the simultaneous assessment of large numbers of genetic markers which are spread throughout the entire genome, in a cost-effective manner by use of high-throughput genotyping microarrays[68]. GWASs enable the identification of disease relevant common genetic variants throughout the entire genome without requiring an a priori hypothesis regarding their location. The hypothesis is that these genetic variants could be anywhere in the genome and that they can be identified by using a systematic strategy. It was the development of catalogues of common variants and their linkage disequilibrium (LD) structures in human populations[69, 70] that enabled the conductance of GWASs. LD is a measure of non-random associations between alleles at different loci and defines the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP[68]. Thus, SNPs on commercial microarrays which are referred to as ‘tagging SNPs’ are selected based on these LD structures and their capability of representing the genetic information from neighbouring markers[68, 71]. For example, 87% of the common variation (minor allele frequency >5%) in

European populations can be captured by using 4.3 million tagging SNPs spread throughout the genome (http://res.illumina.com/documents/products/datasheets/datasheet_omni5.pdf). In a GWAS, each genotyped SNP is tested for its association with the disease status making millions of simultaneously performed statistical comparisons. Therefore; it is necessary to account for laws of probability and adjust the results for eliminating false positive findings. This is achieved by application of stringent correction methods for multiple comparisons that eventually force a quite high genome-wide significance threshold (often taken as $5x10^{-8}$)[72]. GWASs are particularly valuable for diseases with unknown pathogenesis since each implicated gene holds the potential of shedding light onto the biological basis of disease. Up to date GWASs have been performed in several complex human diseases and traits and have generated very valuable information with considerable statistical confidence. Some of these examples include human height, body mass index, metabolic traits and disorders, autoimmune disorders and psychiatric disorders[37, 73, 74].

The first GWASs of schizophrenia have not succeeded in identification of risk alleles surpassing the modern standards for genome-wide significance. This was probably due to relatively small sample sizes that limited power to detect risk alleles of small effect[29]. The first most promising association in schizophrenia GWASs was received in 2008 from an intronic marker in *ZNF804A* gene which yielded an association signal (P=$1.61x10^{-7}$) falling just below the genome-wide significance threshold[75]. Afterwards, the formation of large consortia by the psychiatric genetics community made a significant milestone in genetic research of schizophrenia. In 2009; three back to back major schizophrenia GWAS publications from the SGENE Consortium[76], International Schizophrenia Consortium (ISC)[39] and Molecular Genetics of Schizophrenia Consortium (MGS)[77] reported genome-wide significant association signals implicating major histocompatibility complex (MHC) region (6p21.3-p22.1), neurogranin (*NRGN*) (11q24.2) and transcription factor 4 (*TCF*4) (18q21.2) genes. Genome-wide significance was achieved when independent meta-analyses were performed by exchange of top results among the three consortia. Up to date the most robust common variation association finding for schizophrenia is the MHC region which is highly polymorphic and gene dense harbouring many genes with a wide variety of biological functions including immunity, chromatin modification, transcriptional regulation, neurodevelopment and synaptic plasticity; which are biologically plausible for various aetiological models of schizophrenia[78, 79]. However as the MHC region is characterized by extensive LD, the causal variants and genes in the region are not underpinned up to date. *NRGN* gene is exclusively expressed in brain and particularly in regions important for cognitive functions and in hippocampus important for memory. The gene encodes for a postsynaptic calmodulin (CaM)-binding protein kinase substrate acting as a reservoir making CaM available for activation of postsynaptic calcium/calmodulin-dependent protein kinase II (CaMKII). CaMKII has a major role in synaptic plasticity and memory formation by mediating N-methyl-D-

aspartate (NMDA) receptor signalling[76]. The *TCF4* gene encodes for transcription factor 4 which has an essential role in brain development through its requirement in differentiation of a group of neural progenitor cells[80]. It was shown that mice overexpressing *TCF4* in their forebrains exhibit cognitive impairments as well as pre-pulse inhibition typical of schizophrenia mouse models [81]. In 2011, the previous MHC region, *NRGN* and *TCF4* association signals from SGENE, ISC and MGS studies were replicated in a large follow-up sample (up to 10,260 cases and 23,500 controls) which came up with two novel variants with genome-wide significance; one in the *VRK2* gene encoding for the widely expressed vaccinia related kinase 2 and the other in the vicinity of *TCF4* gene (between coiled-coiled domain containing 68; *CCDC68* and *TCF4*)[82] with an association signal independent from the previously reported variant. In 2011, the previously reported signal from the *ZNF804* gene was followed-up in schizophrenia/schizoaffective disorder (18,945 cases and 38,675 controls) and schizophrenia plus bipolar disorder (21,274 cases and 38,675 controls) sample sets[83] and a signal for the same marker surpassing the genome-wide significance threshold by several orders of magnitude was identified for both sample sets (P=2.5x10^{-11} and P= 4.1x10^{-13}, respectively) which confirmed the initial finding in 2008[75] . Two non-western schizophrenia GWASs have recently been reported in Chinese populations. One of these was the study by Shi *et al.*, who identified two novel loci in 8p12 region implicating *LSM1* and *WHSC1L1* genes and in 1q24.2 region implicating *BRP44* gene[84]. The other study by Yue *et al.*, replicated the MHC region and identified a novel association signal from the 11p11.2 region implicating the *TSPAN18* gene[85]. This region was also reported in another European schizophrenia GWAS which had identified a significant signal in 11p11.2 region implicating *AMBRA1, DGKZ, CHRM4* and *MDK* genes[86] and the signal seemed to be independent from the association signal reported in Yue *et al.*[85]. One the of largest schizophrenia GWASs-also referred to as a mega analysis- was performed by the Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC) which included 21856 individuals of European ancestry in the initial discovery sample and 29839 independent individuals in the replication sample where top-ranked 81 SNPs from the discovery sample were followed-up[87]. The analyses of the initial discovery sample and/or the combined sample set with 51695 individuals identified significant association signals from a total of eight loci. Three of these loci were previously known including the MHC region in 6p21.3-p22.1, 18q21.2 region implicating *CCDC68* and *TCF4* genes, and 11q24.2 region where a signal located ~0.85 megabases (Mb) from *NRGN* gene was identified which was independent from the previously reported *NRGN* signal. Among the five novel loci the strongest signal with a p-value of 1.6x10^{-11}, was received from an intronic SNP located at 1p23.3, within a putative primary transcript of *microRNA 137* (*miR-137*). The other four novel loci were at 2q32.3, 8p23.3, 8q21.3, 10q24.32-q24.33 regions implicating *PCGEM1*, *CSMD1*, *MMP16* and *CNNM2-NT5C2* genes, respectively. The

same study also performed a joint analysis of schizophrenia and bipolar disorder by employment of 16374 bipolar disorder patients and additional 14044 controls. In the joint analysis three loci at 3p21.1, 10q21.2 and 12p13.3 implicating *ITIH3-ITIH4* region, *ANK3* and *CACNA1A* genes, respectively, reached genome wide significance. The strongest finding for schizophrenia in this study, miR-137 is highly expressed in cortex and hippocampus synapses[88] and is a regulator of adult neurogenesis and neuronal maturation which are critical processes in brain development. Moreover, four genes (*TCF4*, *CACNA1A*, *CSMD1* and *C10orf26*) located in regions (18q21.2 , 12p13.3 , 8p23.3 and 10q24.32, respectively) which have reached genome-wide significance in the analysis of schizophrenia and/or joint schizophrenia bipolar samples are verified targets of miR-137[89] suggesting a role for miR-137 mediated dysregulation in etiology of schizophrenia. The most recently published GWAS in schizophrenia describes a multi-stage GWAS analysis which involves i) a large Swedish case-control cohort of 11244 individuals, ii) 20899 individuals from the previously described discovery sample of the PGC mega analysis[87], and iii) independent replication samples (27175 case-control individuals and 581 parent-offspring trios) for replication of SNPs in 168 genomic regions. The analysis revealed 22 genome-wide significant loci of which 13 were novel and 1 was previously implicated in bipolar disorder[90]. Examination of candidate genes at these genome-wide significant loci suggested that neuronal calcium signalling plays a role in the etiology of schizophrenia. The study also generated additional evidence supporting the involvement miR-137 influenced pathways in the disease etiology. Finally the authors estimated 8,300 independent, mostly common SNPs, to contribute to schizophrenia risk which can collectively account for at least 32% of the variance in liability to disease[90].

All together, up to date performed GWASs have been successful in identification of dozens of risk loci harbouring common disease associated variants (majority with a MAF>30%) with small effect sizes (odds ratios mostly <1.2)[29, 90] and every single GWAS signal is valuable in terms of providing some insight into the putative pathophysiological pathways in schizophrenia such as in the case of *miR-137*.

1.3.3 Copy number variation studies

Until recently SNPs were believed to be the major components of variation between individual genomes. Today, we know that submicroscopic structural variants in the form of deleted or duplicated DNA segments are spread throughout the entire genome and they also contribute to genetic variation quite substantially[91-93]. These microdeletions and microduplications ranging from kilobase pairs[90] to megabase pairs in size are referred to as copy number variants (CNVs) which could be common or rare, inherited or *de novo* occurring[93, 94]. Apart from contributing to the normal genetic variation it is established that CNVs are also major mutational events

causing or conferring relatively high risk to a range of neurodevelopmental disorders[95, 96]. A role for structural variants in schizophrenia was already appreciated through some cytogenetic studies as in the case of 22q11.2 microdeletions. However, since recent microarray based technologies enabled the systematic identification and investigation of CNVs in large sample sets, a large body of accumulating evidence has proven a major role for rare CNVs (some of which are *de novo*) in aetiology of schizophrenia as high risk conferring variants. Three main aspects were addressed in schizophrenia CNV studies; the overall burden of inherited and *de novo* CNVs in schizophrenia patients in comparison to controls, and identification of individual CNV loci associated with schizophrenia.

<u>Overall burden of CNVs and *de novo* CNVs in schizophrenia</u>

One of the early studies addressing the burden question in schizophrenia was performed by array comparative genomic hybridization in 150 schizophrenia cases and 268 controls. The authors reported a three-fold increased burden of novel microdeletions and microduplications (>100 kb in size) in schizophrenia cases in comparison to controls[97]. The study showed that the CNVs in cases disrupted disproportionally more often genes implicated in neurodevelopmental pathways. Studies with larger samples followed and one of them was by ISC which performed a genome-wide CNV survey in a sample of 3,391 schizophrenia cases and 3,181 controls by use of high-density SNP microarrays[98]. For rare CNVs larger than 100kb in size and observed in less than 1% in the sample, there was 1.15 fold increased burden in schizophrenia cases and CNVs observed in cases were shown to span 1.41 fold more genes in comparison to CNVs observed in controls. For ultra-rare CNVs observed only in 1 individual a 1.45 fold increased burden in cases, for deletions larger than 500 kb in size a 1.67 fold increased burden in cases and for CNVs observed in cases a 3.57 fold increase in gene count were reported by the authors. A similar increased burden only for gene spanning deletions (>100 kb and >1 Mb) and not duplications was also recently reported by a similar study[45].

Xu *et al.* performed a burden analysis of *de novo* CNVs in 359 trios (152 sporadic cases of schizophrenia, 48 familial cases of schizophrenia, 159 healthy controls and their biological parents)[64] and identified that rare *de novo* CNVs were enriched only in the sporadic schizophrenia cases (~8 times more frequent than in healthy controls, $P = 0.00078$) and not in the familial ones. A complementary scan showed that only in familial cases and not in sporadic ones there was an enrichment of rare inherited CNVs (~2 times more frequent than in healthy controls, $P = 0.01$) which might imply different genetic architectures underlying sporadic and familial cases of schizophrenia[99]. Recently published largest analysis of *de novo* CNVs in schizophrenia (662 schizophrenia cases, 2623 healthy controls and their biological parents) confirmed that *de novo* CNVs were significantly enriched in cases (5.1% of cases vs. 2.2% of controls, P=0.00015) and also reported that *de novo* CNVs in cases disrupted a highly significant

excess of postsynaptic density genes in comparison to the *de novo* CNVs in controls[100] implicating a role for abnormal function of postsynaptic signalling complexes in pathogenesis of schizophrenia.

Identification of specific CNV loci associated with schizophrenia

In 2008, several groups identified heterozygous deletions at 2p16.3 region in schizophrenia patients affecting the *NRXN1* gene [97, 101, 102]. A following study with 2977 schizophrenia patients and 33,746 control subjects established highly significant associations of these microdeletions with schizophrenia[103] which was repeatedly confirmed by following studies[104-107] which made *NRXN1* one of the most robust schizophrenia risk loci[45]. The ISC study[98] from 2008 reported novel specific CNV regions associated with schizophrenia at the 15q13.3 and 1q21.1 regions and confirmed the previously known 22q11.2 microdeletion region. A back to back published paper came from SGENE consortium also reporting these two novel regions and an additional one in 15q11.2 region[108]. SGENE study started with the initial hypothesis of recurrently occurring *de novo* CNVs contributing to stable incidence rates of schizophrenia despite the reduced fecundity associated with disease. The first step was analysis of 9,878 parent-offspring transmissions in a population based discovery sample where they identified 66 *de novo* CNVs which they then tested in 1,433 patients with schizophrenia and related psychoses and 33,250 controls (phase I sample) for association. Three deletions at 1q21.1, 15q11.2 and 15q13.3 regions showed nominal association in the phase I sample, and were followed up in a larger sample of 3,285 cases and 7,951 controls (phase II sample). In the combined sample all three deletions yielded significant association signals with schizophrenia and related psychoses (only 1q21.1 microdeletion was significantly associated with schizophrenia alone). Among the three; 1q21.1 microdeletion had the highest effect size (P=2.9x10^{-5}, OR=14.83) and presented with one short (~1.35 Mb) and one large form (~2.19 Mb) with different putative breakpoints. The segment covered by the shorter form was also common to the longer form and it spanned ten protein coding RefSeq genes of which three fell either completely or partially into the flanking segmental duplications. The presence of these segmental duplications suggested non-allelic homologous recombination to be the likely mechanism responsible for the formation and recurrence of these *de novo* microdeletions[108].

Following these breakthrough CNV findings, further studies established additional microdeletions and microduplications to be genetic risk factors for schizophrenia. Among these are 1q21.1[45], 3q29[109] 7q36.3[45, 109], 16p11.2[110], 16p13.1[111] microduplications and 3q29[45, 112], 17p12[105], 17q12[113] microdeletions. Except for the microdeletions at 2p16.3 and microduplications at 7q36.3 which affect only the *NRXN1* and *VIPR2* genes, respectively, all of the schizophrenia associated CNVs span multiple genes. All together, currently there are several

rare (majority with a control MAF<0.05%) CNVs with intermediate to high effect sizes (odds ratios mostly >5) that are known to be risk factors for schizophrenia[27, 29].

1.3.4 Sequencing studies in schizophrenia

While the large CNVs can be assessed and studied by the use of microarray based SNP data, the role of small rare or low-frequency sequence variants (single nucleotide variants, SNVs; small insertions and deletions, indels) in schizophrenia are addressed by sequencing approaches. The necessity of sequencing arises from the fact that the vast majority of these variants are not tagged in the customized SNP chips and are never likely to be completely catalogued as every individual exome/genome carries hundred/thousands of private variants[114, 115]. The sequencing studies in schizophrenia can mainly be divided between targeted sequencing efforts and exome sequencing efforts which are at their infancy.

1.3.4.1 Region or gene targeted resequencing studies in schizophrenia

Up to date, the vast majority of studies assessing the role of SNVs and indels in schizophrenia relied on targeted sequencing of candidate genes selected on the basis of psychopharmacological hypotheses or some evidence from cytogenetic abnormalities observed in patients and/or linkage analysis findings. The resequencing analysis of vesicular glutamate transporter genes *VGLUT1*[116] and *VGLUT2*[117] based on the glutamatergic hypothesis[16-18] revealed a significant collective overrepresentation of rare variants in schizophrenia patients in comparison to controls. Knight *et al.*, resequenced the exons of the *ABCA13* gene based on their discovery of a schizophrenia patient carrying a complex chromosomal rearrangement disrupting this gene[59]. Genotyping of the multiple rare coding *ABCA13* variants showed their collective frequency to be significantly higher in schizophrenia as well as in bipolar patients compared to controls. The authors suggested a 2.2% and 4.0% population attributable risk for these mutations in schizophrenia and bipolar disorder, respectively. The authors further performed a linkage analysis in families of the mutation carrier individuals and identified co-segregation of the rare variants with psychiatric phenotypes including schizophrenia, bipolar disorder, and major depression. A follow-up study of the *ABCA13* gene from another group; however failed to report similar association findings for rare *ABCA13* variants. In 2008, a study which resequenced *DISC1* gene in 288 schizophrenia patients and 288 healthy controls has identified six patients which were heterozygous carries of 5 rare nonsynoymous variants that were not detected in any of the controls[118]. These variants were moreover absent in a pool of 10000 unrelated control alleles and the authors suggested that ultrarare variants in *DISC1* were associated with an attributed risk of 2% for schizophrenia[118]. A recent study has also identified rare coding *DISC1* variants in 506 cases of schizoaffective spectrum which were absent in 1211 controls[119]. Moens *et al.*[120], performed a large scale mutation analysis of *DISC1* and 10 of its

interaction partners in a case-control cohort ($N_{cases}/N_{controls}$=486/514). The authors stratified their analyses according to functional consequences of variants, minor allele frequencies and the disease onset of the patients and reported a significantly higher burden of rare (MAF <1%) missense variants only in patients with an early age of onset (≤20 age, $P_{corrected}$ = 0.0076). Although *DISC1* locus has not gained any evidence for common variants (i.e. PGC schizophrenia GWASs), the resequencing studies suggest that rare SNVs in the locus might confer susceptibility to schizophrenia. Neverthless, further statistical evidence for rare, *DISC1* and other candidate gene variants are still awaited to survive the modern standards of replication[28].
Recently resequencing studies following-up regions implicated by GWASs are emerging. Shen *et al.*[121], performed resequencing analysis of the *NRGN* gene which is one of the earliest GWAS genes[76, 82] and identified 5 rare variants observed only in patients (6 out of 346 patients) and not in any of the 345 controls. Functional assessment of the identified variants suggested a regulatory effect. Smith *et al.*[122], also performed a mutation screening of *NRGN* in more than 1000 individuals, identified only a singleton non-synonymous variant in the whole study sample and found no evidence for involvement of rare *NRGN* sequence variants in schizophrenia. Dwyer *et al.*, performed a resequencing analysis of the *ZNF804A* gene followed-up by genotyping of the identified rare nonsynonymous variants[123]. The results showed none of the variants to be associated with disease neither when they were assessed individually nor collectively. The exon targeted mutation screening of the *VAV3* gene which yielded one of the top association signals in a Japanese GWAS[124], revealed the presence of 4 rare non-synonymous variants. One of the variants was significantly associated with schizophrenia in the follow-up analysis (*P=0.02,* OR = 0.58) showing a protective effect[125]. The same group performed a similar analysis on two other genes (*KALRN* and *EPHB1*) also implicated by their prior GWAS[124] and reported a significant association of rare missense variants in *KALRN* with schizophrenia both on a collective level and on individual level for one of the analyzed missense variants[126]. No significant associations were found for *EPHB1*.

1.3.4.2 Exome sequencing studies in schizophrenia

As recently advancing technologies are increasing the throughput while reducing the costs of next-generation sequencing[127], there is major interest and promise in application of whole genome or exome sequencing based study designs to schizophrenia. However unlike its applicability and success in rare Mendelian disorders[128-131], this new experimental paradigm faces some major challenges both in research and clinical translation when it comes to complex phenotypes[132-134]. These include the necessity of sequencing large case-control samples (on the orders of 10,000 or more individuals[134]) to distinguish the disease-relevant variants from the background excessive rare variation in the human genome[115, 135] and the ultimate need for

development of comprehensive i) bioinformatical tools for correct annotation and effective prioritization of variants and ii) statistical methods for appropriate handling and interpretation of the vast quantities of data generated. Until now, due to these obstacles the applications of next generation sequencing in schizophrenia were mostly focused on addressing the role of *de novo* mutation in schizophrenia. The benefit of this approach is that less than one *de novo* missense or non-sense mutation is expected to be identified in a parent-offspring trio which is incomparably easier to handle and analyse than the thousands of inherited rare variants in the human exome[136]. Therefore observation of *de novo* mutations affecting the same gene in multiple unrelated affected individuals is very unlikely to occur by chance[137] and when it happens the gene receives substantial support for a likely involvement in the phenotypes etiology[138]. A recent exome sequencing study by Girard *et al.*[139], addressed the general role of *de novo* mutations in schizophrenia by exome sequencing of 14 parent-sporadic proband trios. The authors identified a significantly elevated *de novo* mutation rate in patients (2.59×10^{-8} mutations per base per generation) in comparison to the theoretical expectations of *de novo* mutation rate in humans ($\sim 1.1 \times 10^{-8}$ mutations per base per generation) reported by other studies ($P=0.0025$). Another study from Xu *et al.*[140] published back to back with the Girard *et al.*[139], sequenced the exomes of 53 parent-sporadic proband trios and 22 parent-healthy control trios. The authors did not identify a significant difference between the overall *de novo* mutation rates of the two groups; but reported a significant enrichment and a large excess of putatively functional *de novo* mutations in cases suggesting a substantial contribution of *de novo* protein-altering mutations to sporadic incidences of schizophrenia. The inconsistent findings from Girard *et al.*[139], and Xu *et al.*[140], about the overall rate of *de novo* mutations in sporadic schizophrenia cases can be attributed to sampling effects and methodological differences between the two studies and put emphasis on the value of sequencing matched control samples[141]. Neither of these studies identified genes recurrently affected by mutations in unrelated probands. Xu *et al.*, recently published their latest results from exome sequencing of 231 parent-proband trios enriched for sporadic cases and 34 unaffected trios[138]. The authors again reported similar overall *de novo* mutation rates between patients and controls and also replicated their previous findings by identifying a similar excess of putatively functional *de novo* SNVs as well as a higher prevalence of gene-disruptive (frameshift, nonsense, canonical splice site) *de novo* mutations in cases relative to controls. In this study four genes (*DPYD, TRRAP, VPS39,* and *LAMA2*) were identified to be recurrently affected by two different *de novo* events in unrelated probands and the evaluation of the significance of this finding (P =0.002) suggested it to be very unlikely to have occurred by chance. The evidence from above described exome sequencing studies and others targeting particular regions/genes[142-144] reveal a role for *de novo* variants in schizophrenia. However the theoretical calculations and empirical data suggest that

although the individual effect sizes of *de novo* variants can be large, their overall contribution to the disease liability is most likely to be minor[141]. Therefore the true value of these studies lies in the discovery of novel candidate genes implicating novel pathways or providing additional support to putative risk genes suggested by prior studies. Therefore it is necessary in the near future to extend the exome sequencing based study designs to investigation of the inherited rare and low-frequency variants in the human genome. Up to date only one such study has been published[145]. However in order to overcome the above described challenges in exome sequencing, the authors focused only on investigating if there is a role for the so-called 'goldilocks alleles'[146] which are moderately rare (1-5% MAF) and have moderate to strong effect sizes (2>relative risk≥6) in schizophrenia. The authors conducted a 2-step approach where they exome sequenced a discovery cohort ($N_{cases}/N_{controls}$=166 /307) and followed up the variants they prioritized from the analysis of the discovery dataset (n=5155) by genotyping in an independent cohort ($N_{cases}/N_{controls}$=2617/1800). The authors did not identify any single variants showing a study-wide significant association with disease. The major implications of this study were that there is only a limited role for moderately rare risk variants with strong effects in schizophrenia and that that much rarer variants underlying an allelic heterogeneity are more likely to contribute to disease susceptibility. The authors also pointed out that both large sample sizes and use of gene-based association tests assessing the collective contribution of such rare variants should be necessary to establish significant associations with disease.

Finally a recent exome sequencing study in schizophrenia was reported in multiplex families with several affected individuals[147]. The authors aimed at identifying protein-altering variants co-segregating with disease in these families implicating novel genes and pathways. Such variants were identified in a number of genes involved with glutamatergic neurotransmission in 4 of the 5 families giving some support to the glutamatergic hypothesis in schizophrenia[16-18]. One of the genes (*LRP1B*) was recurrently affected by protein-altering mutations in 3 of the 5 investigated pedigrees.

1.4 The emerging genetic architecture of schizophrenia

Based on the recent evidence from GWASs and CNV studies it is now well acknowledged that the genetic architecture of schizophrenia lies somewhere in between the CDCV and CDRV models. Common variants with modest effect sizes are excluded and completely penetrant mutations are also not expected[27]. At this point, dozens of common SNPs and several rare CNVs have been identified that are associated with schizophrenia having small and relatively large effect sizes, respectively[27, 29]. All these variants are supported by robust statistical evidence. It is expected that substantially larger GWASs will identify many more common risk variants and possibly also further CNVs. Recent exome sequencing studies also suggest a contribution of rare *de novo* point

mutations with possible intermediate to large effect sizes to schizophrenia risk. However none of these individual variants have been replicated or backed up by statistical support in analogy to the CNVs. The same applies to the rare SNVs emerging from candidate gene sequencing studies. Therefore further studies investigating the role of rare and low-frequency sequence variation in schizophrenia are warranted.

1.5 Aims of the doctoral thesis

Based on results from recent systematic, genome-wide association studies in large samples of patients with schizophrenia and controls it has become evident that the genetic variants conferring risk to developing the disease cover a spectrum of disease-associated variants, ranging from rare variants with relatively high penetrance to common variants with individually small genetic effect. It is expected that many more genetic risk variants for schizophrenia await identification. Aim of the present doctoral thesis was to contribute to the understanding of the genetic basis of schizophrenia by identifying more of these unknown susceptibility variants, in particular rare and low-frequency variants. To reach this aim, loci identified previously through either GWASs or CNV studies were planned to be resequenced with the underlying rationale that different types of genetic variations can contribute to disease susceptibility at an individual locus[148, 149]. A resequencing based genetic screening approach would be excepted to provide an overview of rare and low-frequency variants in that particular locus which fall below the resolution of array-based technologies. This information can subsequently be used to i) assess the contribution of rare and low-frequency variants to disease susceptibility, ii) define which parts of an associated region are responsible for the observed association and iii) identify potentially deleterious variants which can be subjected to functional assays to obtain insight into pathophysiology[148].

Specifically, the goal of the current thesis was i) to perform targeted sequencing in 3 different regions robustly associated with schizophrenia by multiple GWASs or CNV studies in schizophrenia in order to get a comprehensive picture of the low-frequency and rare sequence variation present at these loci; ii) to investigate the association of these low-frequency and rare sequence variants with schizophrenia by genotyping in large samples of patients and controls, and iii) to deliver schizophrenia-associated variants that can be subjected to functional studies in future for better understanding of the disease pathophysiology.

Among the three regions studied were two regions that had emerged from CNV studies, the *NRXN1* gene and the 1q21.1 microdeletion region, and one region that had emerged from GWASs, the *TCF4* gene. In all of the three regions, discovery samples comprised of about 190 individuals and follow-up samples ranging from about 2500 to 9000 individuals were used. Discovery samples were resequenced for the region of interest to discover rare and low-

frequency variants and follow-up samples were used for genotyping of the identified variants. Various statistical tests were performed to test for the association of rare and low-frequency variants in each of these regions with schizophrenia. Complementary analyses like expression and splice site analyses were performed when applicable.

The results of the *NRXN1* study have already been published[150] whereas the *TCF4* study has recently been submitted (Basmanav *et al*., submitted to *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*) and the study on 1q21.1 microdeletion region is in preparation for submission (Basmanav *et al.,* in preparation for submission to *Schizophrenia Research)*

2. MATERIALS AND METHODS

2.1 Devices

Autoclaves
- Systec D-150, Systec GmbH
- Varioklav® 135 S Dampfsterilisator, H+P Labortechnik GmbH

Centrifuges
- Megafuge 1.0 R, Heraeus GmbH
- Biofuge fresco, Heraeus GmbH
- Biofuge pico, Heraeus GmbH
- neoLab Mini-Centrifuge Spectrafuge®, neoLab
- Concentrator Plus System, Eppendorf AG

Concentration measurement devices
- BioAnalyzer 2100, Agilent Technologies Deutschland GmbH
- NanoDrop® 1000 Spectrophotometer, Peqlab Biotechnology GmbH
- NanoDrop® 8000 Spectrophotometer, Peqlab Biotechnology GmbH

DNA storage systems
- 2D CYPHER Tubes, Thermo Fisher Scientific GmbH
- 2D CYPHER 1,2ml Cluster Tube Racks, Thermo Fisher Scientific GmbH
- SmartScan Solo™ 2D Barcode Reader, Thermo Fisher Scientific GmbH
- SmartScan 96 2D Barcode Reader, Thermo Fisher Scientific GmbH

Drying chamber
- T 20 P, Heraeus GmbH

Electrophoresis chambers
- WIDE MINI SUB CELL® GT, BioRad Laboratories GmbH
- Sub-Cell Model 96, BioRad Laboratories GmbH

Gel documentation
- GelDoc™ XR System, BioRad Laboratories GmbH

Genotyping systems
- iScan System, Illumina® Inc.
- MassARRAY™ Compact Analyzer, Bruker Daltonics Inc. for Sequenom® GmbH

Ice machine
- AF100, Scotsman® Ice Systems

Isolation of nucleic acids
- Magnetic Separation Module I, Perkin Elmer Chemagen Technologie GmbH

Mix and stir devices
- REAX 2 / TITRAMAX 101 / UNIMAX 1010, Heidolph Instruments GmbH & Co. KG
- Thermomixer comfort, Eppendorf AG
- Vortex Genie 2, Scientific Industries Inc.
- Vortex Mixer IKA MS2-S8, Agilent Technologies Deutschland GmbH

Nanodispenser
- MassARRAY™ Nanodispenser, SAMSUNG Techwin Co. Ltd. for Sequenom® GmbH

Pipettes
- Research® variable pipette set (0.1-2.5 µl, 0.5-10 µl, 2.0-20 µl, 10-100 µl, 20-200 µl,(100-1000 µl, 500-5000 µl), Eppendorf GmbH
- Eight channels pipettes (0.5-10 µl, 10-100 µl), Eppendorf GmbH
- Finnpipette® 16 channels, VWR International GmbH
- Multipette® plus, Eppendorf AG
- Transferpette®, BRAND GmbH & Co. KG
- Transferpette® S-8, BRAND GmbH & Co. KG
- Transferpette®-8/-12 electronic, BRAND GmbH & Co. KG

Pipette robot systems
- Biomek® Laboratory Automation Workstations NX MC and NX S8G, Beckman Coulter GmbH

Pipette support
- accu-jet® pro, BRAND GmbH & Co. KG
- Cell Mate II, Matrix Technologies Corporation, Thermo Fisher Scientific Inc.
- Chip Priming Station, Agilent Technologies Deutschland GmbH

Power supply
- PowerPac Power Supplies, Bio-Rad Laboratories GmbH

Scales
- TE3102S / TE3135-DS, Sartorius AG

Sequencing device
- 3130xl Genetic Analyzer, Life Technologies GmbH

Sterile hood -
- HERA*safe*, Heraeus GmbH

Thermal cycler
- PTC-200 and PTC-100, MJ Research Inc.
- ABI Prism® 7900HT Fast Real-Time PCR System (TaqMan), Life Technologies GmbH

Water purification system
- Milli-Q A10 Synthesis, Merck KGaA

2.2 Chemicals, Reagents, Buffers and Solutions

- Agarose low EEO, AppliChem GmbH
- Alconox, Alconox Inc.
- Bromphenol Blue, Sigma-Aldrich Chemie GmbH
- Dimethyl sulfoxide; DMSO (C2H6SO), Sigma-Aldrich Chemie GmbH
- DNase I, Qiagen GmbH
- dNTPs [10mM], Labomedic GmbH
- Ethanol absolut; EtOH (C2H5OH), AppliChem GmbH
- EtOH 96%, WALTER CMP GmbH & Co. KG
- Ethidium Bromide; EtBr (C21H20N3Br) 1%, Merck KGaA
- Fetal calf serum (FCS), Biochrom AG
- Ficoll 400, GE Healthcare GmbH
- Fungizone (250 µg/ml amphotericin B, 205 µg/ml sodium deoxycholate), Invitrogen Co.
- HotStar Taq DNA Polymerase [5 U/µl], Qiagen GmbH
- HPLC water, Merck KGaA
- Isopropanol (C3H8O), AppliChem GmbH
- L-Glutamine [200 mM], Biochrom AG
- Loading buffer: 10ml 10X TBE, 10ml 0,1%Bromphenol Blue, 40ml 20% Ficoll, 40ml dH_2O
- Magnesium chloride (MgCl2) [25mM], Qiagen GmbH
- PCR Buffer (10X) with MgCl2 [15 mM], Qiagen GmbH
- Penicillin/Streptomycin (Pen/Strep) (100 x), Invitrogen Co.
- Phytohemagglutinin L (PHA-L), Biochrom AG
- Proteinase K, Qiagen GmbH
- RNaseZap™-Solution, Ambion®, Life Technologies GmbH
- RPMI 1640 (2.0 g/l NaHCO3, without L-Glutamine), Biochrom AG
- Sodium hypochlorite (NaOCl) (13%), AppliChem GmbH
- Sodium pyruvate, Life Technologies GmbH
- 0,4% Sodium hypochlorite solution: 15,4ml NaOCl, 484,6ml dH_2O
- 10X TBE Buffer, Life Technologies GmbH
- Tris-BASE (NH2C(CH2OH)3) (Trizma), Sigma Aldrich Chemie GmbH
- Tris-Cl [10mM], pH 8,0: 0,6g Tris-BASE, 500ml dH_2O.
- Tris-EDTA (TE-4): 0,1mM EDTA, 10mM Tris-Cl pH 8,0

- VWR-*Taq* DNA-Polymerase [5U/µl], VWR International GmbH

2.3 Commercial systems

- 5X Big Dye Terminator Cycle sequencing Kit 3.1, Life Technologies GmbH
- AMPureKit, Agencourt Bioscience Corp.
- AllPrep™ DNA/RNA Micro Kit, Qiagen GmbH
- Chemagic DNA Blood Kit special, Perkin Elmer Chemagen Technologie GmbH
- CleanSEQ Kit, Agencourt Bioscience Co.
- DNA Ladder (100bp, 1kb) AppliChem GmbH
- DNA Ladder 100bp, New England Biolabs Inc.
- Illumina® HT12-v3 Expression BeadChip Kit, Illumina® Inc.
- Illumina®TotalPrep™-96 RNA Amplification Kit, Ambion, Life Technologies GmbH
- iPLEX™ Gold Reagent Kit, Sequenom® GmbH
- RNA 6000 nano LabChip Kit, Agilent Technologies Deutschland GmbH
- RNase-free DNase Set, Qiagen GmbH
- RNeasy Micro Kit, Qiagen GmbH
- SpectroCHIP® Arrays & Clean Resin Kit, Sequenom® GmbH
- SuperScript III First-Strand Synthesis System for RT-PCR, Life Technologies GmbH
- TaqMan® Gene Expression Master Mix, Life Technologies GmbH
- TaqMan® Gene Expression Assay "Hs00188720_m1" (*CHD1L*), Life Technologies GmbH
- TaqMan® Endogenous Control Assay Human Cyc "4326316E" (*Cyclophilin*), Life Technologies GmbH

2.4 Softwares and databases

- 1000Genomes Browser (http://browser.1000genomes.org/index.html)
- 2D CYPHER™ Pilot Databases, Thermo Fisher Scientific GmbH
- 3130xl DataCollection v3.0, Life Technologies GmbH
- Allen Brain Atlas (http://www.brain-map.org)
- Assay Design 3.1, Sequenom® GmbH
- BeadScan, Illumina® Inc.
- Bioconductor (www.bioconductor.org)
- Biomek® Software 3.2, Beckman-Coulter GmbH
- BrainSpan Atlas (http://www.brainspan.org/lcm/search/index.html)
- DNA Calculator (http://www.sigmaaldrich.com/life-science/custom-oligos/custom-dna/learning-center/calculator.html)
- DomPred (http://bioinf.cs.ucl.ac.uk/dompred)
- ENSEMBL genome browser (http://www.ensembl.org/index.html)
- Expasy Prosite (http://prosite.expasy.org/)
- Expasy Translate (web.expasy.org/translate/)
- GenomeStudio™ v2011.1, Illumina® Inc.
- GWAS online catalog (http://www.genome.gov/gwastudies)
- HapMap genome browser (http://hapmap.ncbi.nlm.nih.gov)
- Human Splicing Finder (http://www.umd.be/HSF/)
- Jpred3 (http://www.compbio.dundee.ac.uk/www-jpred/)
- Mutation taster (http://mutationtaster.org/)
- MySequenom, Assay Design Suite (https://mysequenom.com/Home)
- NanoDrop® ND-100 v3.3.0, Peqlab Biotechnology GmbH
- NanoDrop® ND-8000 v2.2.1, Peqlab Biotechnology GmbH
- NCBI (http://www.ncbi.nlm.nih.gov)
- OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)
- Polyphen (http://genetics.bwh.harvard.edu/pph2/)

- Pubmed (http://www.ncbi.nlm.nih.gov/pubmed/)
- PRINTS database (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php)
- R-Version 2.15.2 (http://www.r-project.org)
- RT-Workstation 3.3, Sequenom® GmbH
- SDS 2.2.2, Life Technologies GmbH
- SeqMan II Version 5.0, DNASTAR Inc.
- SIFT (http://sift.jcvi.org/)
- SMART (http://smart.embl-heidelberg.de/)
- SmartScan™ 96, Thermo Fisher Scientific GmbH
- SpectroAQUIRE, Version 3.3.1.2, Sequenom® GmbH
- SpectroPOINT, Sequenom® GmbH
- Typer 3.4/ 4.0, Sequenom® GmbH
- Uniprot (www.uniprot.org/)
- UCSC (http://genome.ucsc.edu/)

2.5 Study Samples

In this thesis 3 resequencing projects are described each having different work-flows. Cohorts of schizophrenia patients and healthy control individuals ($n_{total=}$9055) have been recruited in Germany (n=4281), Denmark (n=3717) and the Netherlands (n=1057).
German patients who were diagnosed according to DSM-IV criteria for schizophrenia ($n_{total=}$2001) were recruited at i) Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Germany (n=842), ii) Department of Psychiatry, Ludwig-Maximilians-University, Munich, Germany (n=930) and iii) Department of Psychiatry, University of Bonn, Germany (n=229). Population based German controls ($n_{total=}$2280) were screened for schizophrenia and not for other psychiatric phenotypes. A part of the controls were drawn (n=1113) from a large cohort collected in the context of the Heinz Nixdorf Recall Study, Germany[151] and the remaining (n=1167) were collected at the Central Institute of Mental Health,Mannheim, Germany.
DNA samples from schizophrenia patients diagnosed according to DSM-IV criteria (n_{total}=646) and population based controls (unscreened for schizophrenia and other psyhiatric phenotypes) (n_{total}=411) with Dutch origin were provided by the University Medical Center Utrecht, Department of Medical Genetics in Utrecht, Netherlands.
DNA samples from schizophrenia patients diagnosed according to ICD-10 (1994–2005) criteria (n_{total}=1871) and population based controls (n_{total}=1846) with Danish origin were provided by the Statens Serum Institute, Section of Neonatal Screening and Hormones in Copenhagen, Denmark. None of the Danish controls had been assigned a diagnosis of schizophrenia at the time of recruitment according to Danish health registers[152].
Different combinations of individuals from these samples have been used in different steps of each project as described below.

NRXN1 project

Resequencing sample for the *NRXN1* gene consisted of 94 schizophrenia patients (49 male/45 female) and 94 sex-matched controls (49 male/45 female) of German origin. The patients were drawn from the schizophrenia cohort recruited at the Central Institute of Mental Health, Mannheim and at the Department of Psychiatry of the University of Bonn, Germany, based on the criteria that parental DNA and genome wide genotype data were available to check for *de novo* events and to exclude the presence of copy number variants at the *NRXN1* locus, respectively. Compatible with these criteria were only 89 patients derived from parent-offspring trios. For the remaining 5 patients, parental DNA was available only from one parent. Thereby additional selection criteria, namely an early age of onset (≤21 years of age; defined by first occurrence of symptoms) and/or positive family history of schizophrenia (defined by affection of at least one first-degree relative) were used for these patients to increase the likelihood for identification of highly penetrant variants. The controls were a small subset of the population based control cohort recruited within the context of the Heinz Nixdorf Recall Study[151]. The mean age-at-recruitment for the controls was 47.9 years. All of the resequenced individuals were screened for the absence of copy number variants at the *NRXN1* locus (chr2:50002456-51109064, NCBI build 36.1) based on their SNP intensity data from Illumina HumanHap550v3 and HumanHap610v1 BeadArrays using the QuantiSNP v1.1[153] software.
For the follow-up genotyping step a subset of the total German cohort was used which consisted of 1415 schizophrenia patients (818 male/597 female) and 1167 controls (569 male /598 female).

1q21.1 microdeletion region project

1) The same resequencing sample (94 patients and 94 controls of German origin) defined in the *NRXN1* project was used for the resequencing of the 7 genes in the 1q21.1 microdeletion region. This sample will be referred to as the "initial sequencing sample" in the context of this study. All of the individuals in this sample were screened for the absence of copy number variants at the 1q21.1 microdeletion region (chr1:144943150-146293282, NCBI build 36.1) based on their SNP intensity data from Illumina HumanHap550v3 and HumanHap610v1 BeadArrays using the QuantiSNP v1.1[153] software. In order to assess a potential unmasking of recessive variants, two female patients who were known to be 1q21.1 microdeletion carriers were additionally sequenced for the 7 genes. One of these patients was previously desribed[108].

2) A genotyping based follow-up step was performed using an independent German case-control sample composed of 1900 schizophrenia patients (1142 male/758 female) and 2186 control individuals (1056 male/1130 female). This sample will be referred to as "Genotyping-1 sample".

3) A secondary sequencing sample of 96 German patients (49 male/47 female) was used for extended sequencing of two prioritized genes in the region. This sample will be referred to as the "extended sequencing sample". The extended sequencing sample was a part of the genotyping-1 sample and was drawn from the German schizophrenia cohort based on the presence of a positive family history of schizophrenia to increase the likelihood for identification of highly penetrant variants. The extended sequencing sample was also screened for the absence of copy number variants at the 1q21.1 microdeletion region.

4) Another genotyping based follow-up step was performed using i) a German case-control sample composed of 1808 schizophrenia patients (1094 male/714 female) and 2186 control individuals (1056 male/1130 female). This sample represented the total of the available German case-control cohort excluding the initial and the extended sequencing samples and thus had a major overlap (~95%) with Genotyping-1 sample, ii) the Dutch cohort of 646 schizophrenia patients (477 male/169 female) and 411 control individuals (213 male/198 female) and iii) the Danish cohort of 1871 schizophrenia patients (1059 male/812 female) and 1846 control individuals (1043 male/803 female). All together this genotyping sample consisted of 8768 individuals and will be referred to as "Genotyping-2 sample".

TCF4 project

The resequencing efforts in the *TCF4* gene were focused only on patients. A total of 190 schizophrenia patients (96 male/94 female), majority of which (n=185) overlapped with the initial and extended resequencing samples of the 1q21.1 microdeletion region, were selected for resequencing of the *TCF4* gene. The non-overlapping five patients were selected for a positive family history of schizophrenia from the German patient cohort. The variation in control populations in this region was accounted for by using the rare variant information from the 379 European individuals of the 1000 Genomes Project. The European individuals defined in the 1000 Genomes project emerge from 5 sub-populations which include Utah residents with Northern and Western European ancestry (CEU) (n=85), Toscani in Italia (TSI) (n=98), British from England and Scotland (GBR) (n=89), Finnish from Finland (FIN) (n=93) and Iberian populations in Spain (IBS) (n=14)[154].

For the genotyping step in this project a German case-control cohort composed of additional 1808 schizophrenia patients (1095 male/713 female) and 2261 healthy control individuals (1096 male/1165 female) was used.

To follow-up the molecular genetic findings two sample sets were used for functional analyses.

Epstein–Barr virus (EBV)-transformed lymphoblastoid cell lines

EBV-transformed lymphoblastoid cell lines from 20 individuals (7 males/13 females, mean $age_{recruitment}$=45.75) were used for RNA isolation and allele-specific expression analysis. These

individuals were drawn from a cohort of 280 individuals based on their genotype at rs36008075 located in the *CHD1L* gene in the 1q21.1 microdeletion region. The genotypes were determined by sequencing analysis of the whole cohort. The cohort consisted of Spanish individuals from whom both DNA samples and EBV-transformed lymphoblastoid cell lines were available. The individuals were diagnosed with bipolar spectrum disorder acccording to DSM-IV criteria.

Pre-mortem hippocampus tissue samples

Biopsy samples from patients with chronic pharmacoresistant temporal lobe epilepsy (n=148) were collected in the Epilepsy Surgery Program at Bonn University. Surgical removal was necessary for seizure control in all patients after standardized presurgical evaluation by combination of noninvasive and invasive procedures[155]. Fresh frozen pre-mortem human hippocampal segments were provided by the Bonn Tissue Bank. These samples were collected and processed in collaboration with Prof. Dr. med. Albert Becker from University of Bonn.

The studies presented in this thesis were approved by the ethics committees of all study centers. Each participant provided written informed consent prior to inclusion, and all aspects of the study complied with the Declaration of Helsinki.

2.6 Methods

2.6.1 Isolation and management of nucleic acids

Isolation of DNA from blood

Total human DNA from lymphocytes was isolated using either salting-out method[156] or a Chemagic Magnetic Separation Module I (Figure 2.1 a) and chemagic DNA blood kits according to the manufacturer's instructions (http://www.chemagen.com/). The salting-out method relies on the initial lysis of the cells followed by Proteinase K digestion and precipitation of DNA with isopropanol.

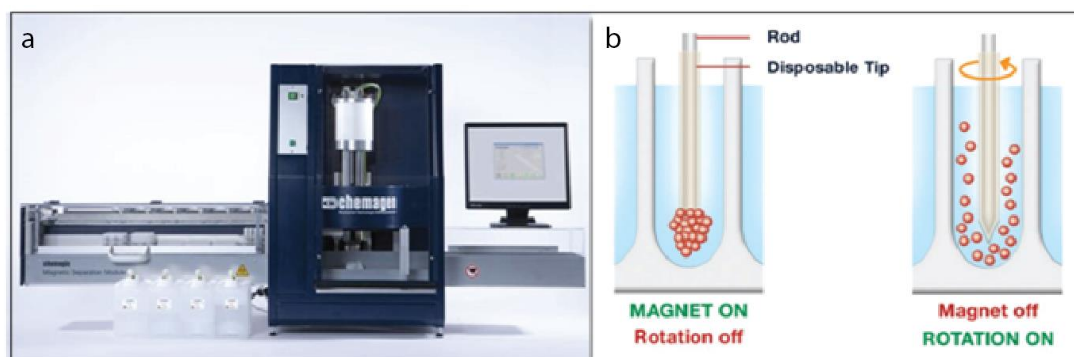


Figure 2.1 Isolation of DNA from blood samples. (a) Chemagic Magnetic Separation Module I (Taken from http://www.chemagen.com), (b) Interaction of DNA-bead complexes with magnetized and de-magnetized metal rods (Taken from http://www.abbis.de).

The isolation by Chemagen relies on binding of the DNA molecules to magnetic poly-vinyl alcohol beads (M-PVA magnetic beads). Under an applied electromagnetic field; DNA-bound magnetic beads are attracted to metal rods which can transfer the DNA-bead complexes to several washing buffers. At the end of each transfer step, the electromagnetic field is switched off and magnetic rods start to rotate leading to re-suspension of the DNA-bead complexes (Figure 2.1 b). The final transfer is made into the elution buffer which elutes the DNA molecules from the magnetic beads.

The native DNA stock solutions were stored at -80°C. The working dilutions of 100ng/µl and 20ng/µl were prepared using the Abgene DNA storage system and stored at -20°C and 4°C for long or short-term storage, respectively. DNA samples were handled at room temperature (RT) during conductance of experiments as DNA has a relatively stable structure with only little susceptibility to spontaneous degradation by catalytic hydrolysis. All the pipette tips and reaction tubes were autoclaved prior to use to avoid contamination.

Isolation of total RNA from immortalized cell lines

RNA was isolated from EBV-transformed human lymphoblastoid cell lines which were grown in RPMI 1640 medium containing fetal calf serum under optimum culture conditions (37^{0}C and 5% CO_2). Penicillin, streptomycin and amphotericin B were added into the culture medium to avoid bacterial and fungal contamination. RNA isolation was performed by use of the RNeasy Micro Kit according to manufacturer's protocol (http://www.qiagen.com/).

RNA samples were only stored at -80°C and were always handled on ice during conductance of experiments due to the more sensitive structure of single-stranded RNA. In order to avoid their spontaneous degradation by ubiquitously abundant RNAses[157] all the equipments and the bench tops used during conductance of RNA experiments were kept RNAse-free by use of commercially available RNaseZap™ solution. RNAse free water was provided within commercial kits. All the pipette tips and reaction tubes were commercially provided as RNAse-free.

Isolation of DNA and RNA from hippocampus tissue

Total DNA and RNA were extracted from fresh frozen pre-mortem human hippocampal segments provided as tissue-slices prepared via cryostat-conditions. The isolations were performed by using the AllPrep DNA/RNA Micro Kit following the manufacturer's protocol (http://www.qiagen.com/).

Quality control and quantification

The concentration of DNA and RNA samples were measured by NanoDrop-1000 or NanoDrop-8000. The spectrophotometric concentration measurement relies on the linear correlation between the amount of the absorbed light and the concentration of the absorbing

molecule predicted by the Beer Lambert law. The maximal absorption wavelength of nucleic acids is 260 nm. When the DNA or RNA samples are exposed to ultraviolet light at this wavelength they will absorb it and less light will be detected by the photodetector yielding a higher optical density (OD). One unit of OD corresponds to 50 ng/µl and 40 ng/µl of DNA and RNA, respectively. Proteins absorb light at 280 nm. The purity of samples is assessed by measurement of the ratio of absorbance at 260 and 280 nm (A260/280). An A260/280 value of 1.8 and 2.0 were taken as the standards to validate the purity of DNA and RNA samples, respectively.

2.6.2 Reverse transcription

Total RNA samples were used for reverse transcription of messenger RNA (mRNA) to complementary DNA (cDNA) by using the 'SuperScript III First-Strand Synthesis System for RT-PCR' kit. Targeted reverse transcription of mRNA population from the total RNA relies on specific hybridization of oligo(dT) molecules to the poly(A)-tail of mRNA molecules. This is followed by the synthesis of a cDNA strand with a retroviral reverse transcriptase which uses the Oligo(dT) primed RNA strand as the template. The following protocol was used for reverse transcription:

1. The following RNA/primer mixture adding up to a total volume of 10 µl is prepared for each sample and incubated at 65°C for 5 minutes.
    - 1 µl oligo(dT) [50 µM]
    - 1 µl dNTP mix [10 mM]
    - 300 ng (x µl) total RNA
    - RNAse free water (8-x µl)
2. The samples are placed on ice for at least 1 minute and the following cDNA synthesis mix is prepared for the total number of reactions to be performed.

Components	Volume per reaction (µl)
10x RT PCR buffer	2
$MgCl_2$ [25 mM]	4
DTT [0.1 M]	2
RNaseOUTTM [40U/µl]	1
SuperScriptTM III RT [200U/µl]	1

3. Each RNA/primer mixture is filled up with 10 µl of cDNA synthesis mix and incubated at 50°C for 50 minutes for the cDNA synthesis to take place.
4. Reactions are terminated by incubation at 85°C for 5 minutes and the tubes are chilled on ice.

5. For degradation of the residual RNA, 1 µl of RNase H is added to each tube and incubated at 37°C for 20 minutes.

The cDNAs were either stored at -20 °C or used immediately.

2.6.3 Real-Time Quantitative PCR

Real-Time quantitative PCR (RT-qPCR) is a method used for quantification of gene expression. The method uses cDNA reverse transcribed from mRNA. However, in contrast to a normal PCR, the RT-qPCR allows real-time quantification of the DNA product at the end of every PCR cycle as the amplification reaction is ongoing. RT-qPCR can be applied for absolute quantification of the input template or its quantification relative to a reference sample. The method relies on quantification of fluorescence signal intensity which correlates with the amount of amplified product. For this purpose different fluorophores can be used either simultaneously with or attached to transcript specific probes. TaqMan probe–based chemistry is one of the conventional examples of the latter approach. A Taqman assay relies on the fluorescence resonance energy transfer (FRET)[158] and the 5'-3' exonuclease activity of the Taq Polymerase enzyme (Figure 2.2) and it can be referred to as a 5' nuclease assay. A transcript specific hybridization probe which is linked to a fluorescent reporter dye at its 5' end and a non-fluorescent quencher at its 3' end is the essential component of the 5' nuclease assays (Figure 2.2 a). When the probe is intact and these two molecules are in close proximity, all the energy emitted from the florescent dye upon excitation is absorbed by the quencher molecule according to the FRET principle and no fluorescence signal can be detected. As the Taq Polymerase encounters the probe during the strand extension process, it performs strand displacement and cleaves the reporter dye by using its 5'-3' exonuclease activity (Figure 2.2 b, c). The cleavage of the reporter dye enables the release and detection of the fluorescence. The intensity of the detected signal is proportional to the amount of template present in the reaction. The polymerization of the strand continues as the rest of the probe is fragmented (Figure 2.2 d). The quantification is made by determination of a cycle threshold value (C_T) for each sample which specifies the cycle number at which the amplification process is in an exponential phase and the generated fluorescence signal exceeds an automatically set detection threshold value common for all samples. In accordance, the higher the template abundance, the sooner (at an earlier cycle) a significant increase in fluorescence surpassing this threshold value will be observed. In order to correct for possible variation in input amounts and quality of templates across different samples, an assay targeting a reference gene, also referred to as an endogenous control is run simultaneously and the C_T data from the endogenous control is used to normalize the quantitative data from the target gene. In this study quantitative Real-Time PCR was used for relative gene expression analysis. The aim was to determine allele specific

expressions of *CHD1L* gene at rs36008075. RT-qPCR was performed by using a Taqman assay which targeted the *CHD1L* gene (Hs00188720_m1). The reactions for each sample were run in quadruplicates with 2 µl of cDNA as template input.

The reference gene used in this study was *cyclophilin A* and the assay (4326316E) was also commercially available from Applied Biosystems.

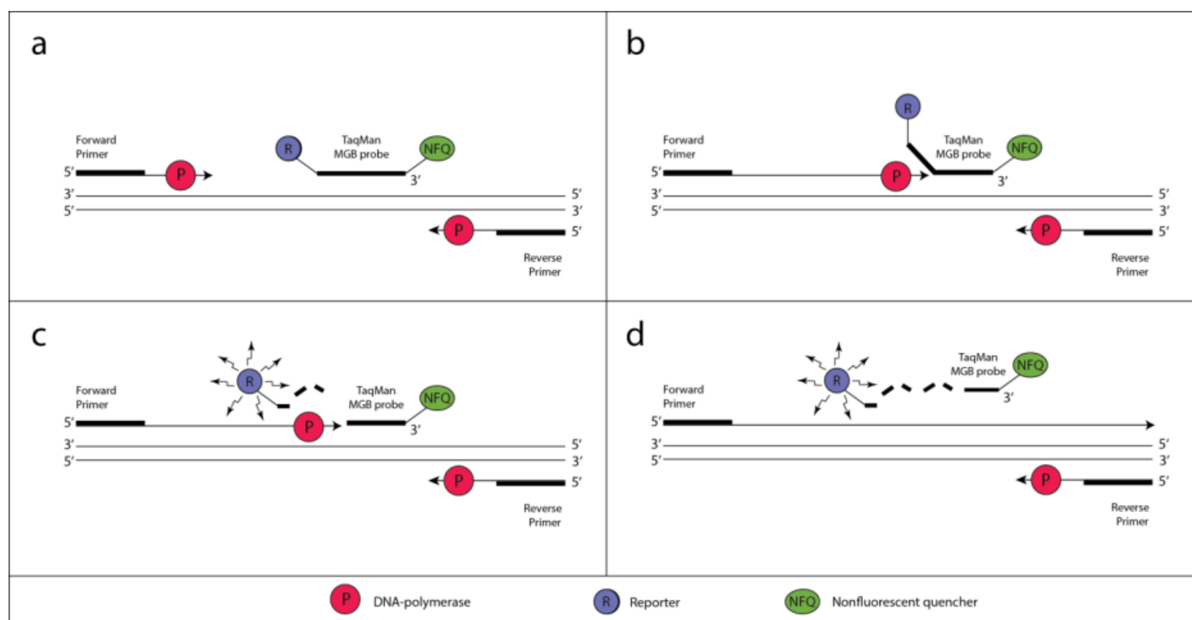


Figure 2.2 Schematic representation of the TaqMan® 5' nuclease assay chemistry. (a) Hybridization of the probe to the target region and initiation of polymerization (b) Strand displacement (c) Cleavage of the reporter molecule and fluorescence emmision (d) Completion of the polymerization. Created based on the illustration from Taqman® Gene Expression Assays Protocol. (http://tools.invitrogen.com/)

The reactions were run in the Applied Biosystems 7900HT Fast Real-Time PCR System. The C_T data was analyzed according to the $\Delta\Delta C_T$ method[159] and the expression levels were determined relative to the mean expression level of individuals homozygous for the wild type allele which was set to 100%. The following tables show the reaction mixture for a single reaction adding up to a total volume of 10 µl and the cycle set up, respectively.

Components	**Volume per reaction (µl)**
2x TaqMan® Gene Expression Master Mix	5
20x TaqMan® Gene Expression Assay (for target transcript)	0.5
20x TaqMan® Gene Expression Assay (for endogenous control)	0.5
dH_2O	2
Template (cDNA)	2

10 min	95°C	Initial denaturation	
15 sec	95°C	Primer annealing and extension	40 cycles
1 min	60°C		

2.6.4 Sanger sequencing

Sanger sequencing is a method for defining the base-by-base sequence information of a genomic region by using dideoxynucleotide triphosphate (ddNTP) chain termination[160].

The method is mainly comprised of the following steps. Exponential amplification of the target region is the first step to enable extraction of the specific information of the targeted region from the whole genome and polymerase chain reaction (PCR) is the standard tool for this purpose[161]. For this reaction a heat stable DNA polymerase, deoxynucleotide triphosphates (dNTPs) and primers pairs designed specifically for the region of interest are used. A PCR functions by cyclic temperature changes allowing 1) denaturation of the double stranded DNA, 2) annealing of the primer pairs to the target sequence and 3) DNA synthesis (elongation) in a cyclic manner. In order to confirm successful amplification of a target region, standard agarose gel electrophoresis is performed after each PCR. The main principle is the migration of negatively charged DNA molecules through a porous agarose matrix under the influence of an applied electric field. The products are controlled for their expected sizes with reference to DNA ladders co-loaded into the gel. Fluorescent dyes which can intercalate with nucleic acids are used to visualize DNA bands by fluorescing under ultraviolet light. Once the products are verified, they are cleaned from residual components of the PCR reaction by a cleaning step and then are subjected to cycle sequencing reaction. Cycle sequencing is similar to the first PCR step with regards to cyclic temperature changes moderating denaturation, primer annealing and elongation steps. However, in contrast only one primer molecule (forward or reverse) is used. In a cycle sequencing reaction, besides the normal dNTPs, fluorescently labeled ddNTPs are included in the reaction mixture[160]. Each of the ddNTP type (ddATP, ddTTP, ddCTP, ddGTP) is marked with a different fluorescent dye. As the elongation is ongoing, ddNTPs compete with the dNTPs. As soon as a ddNTP is incorporated into an extending DNA fragment, the reaction is terminated. Eventually when the reaction is over, DNA fragments of different lengths encompassing all possible fragment sizes in the targeted region are obtained, each of them ending with a fluorescently labeled ddNTP (Figure 2.3 a). After the cycle sequencing products are purified from residual components by a cleaning step they are subjected to capillary electrophoresis. The principle of capillary electrophoresis is the migration of DNA fragments through polymer filled capillaries in different speeds due to their different sizes. As fragments reach the end of the capillary from short to longer ones consecutively, they are excited by a laser beam and generate fluorescent signals which are detected and transformed into electropherograms yielding base-by-base sequence information (Figure 2.3 b).

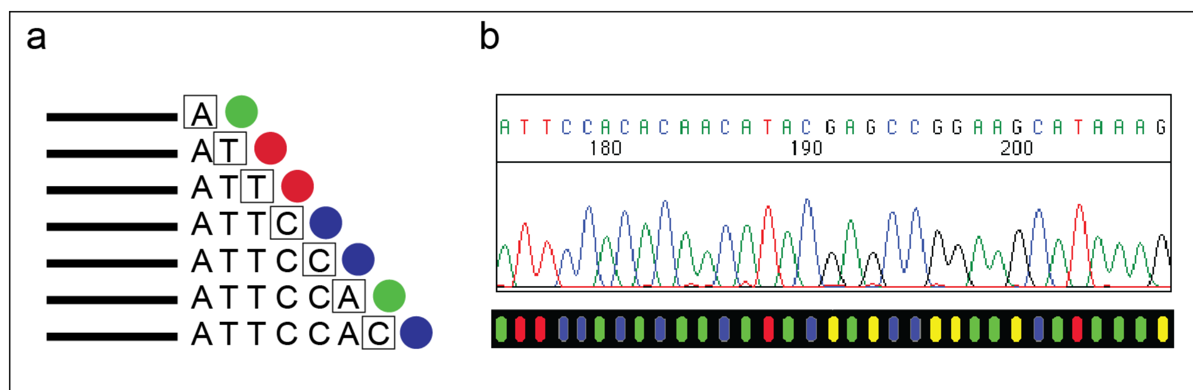


Figure 2.3 Cycle sequencing and capillary electrophoresis (a) DNA fragments of different lengths each ending with a base-specific fluorescent dye are obtained at the end of the cycle sequencing reaction (b) The products from cycle sequencing reaction are injected into a single capillary and are subjected to electrophoresis (bottom). As products of different lengths reach th e end of the capillary at different time points, they are excited by a laser beam and generate fluorescent signals which are detected and transformed into electropherograms (top). Taken from Applied Biosystems Chemistry Guide, Second edition (tools.invitrogen.com/content/sfs/manuals/cms_041003.pdf)

Primer design

Genomic DNA sequences of the targeted genes were obtained from the UCSC Genome Browser based on either NCBI 36.1/hg18 (*NRXN1* gene and 1q21. microdeletion region) or GRCh37/hg19 (TCF4 gene) assemblies. Targeted regions were the exons and the flanking up- and downstream splice site sequences. A list of all targeted transcripts as well as primer pairs and the lenght of the amplicons they generate are provided in the Attachment I and III, respectively. Between each primer and the target sequence a minimal distance of 50 bp was maintained to guarantee high electropherogram qualities. The specificity of primers was validated against the human reference genome using BLAT tool in the UCSC Genome Browser. The anneling temperatures of the primer pairs were in the range of 55-68^{0}C and the allowed maximum difference between the annealing temperatures of reverse and forward primers of each primer pair was 3^{0}C. The GC content of primers ranged between 40-60%. All these criteria were controlled by the online tool DNA calculator.

Amplification

The master mixture -with a 25 µl of total volume/per reaction- and the cycle set up used in the initial PCR are given in the following tables. To increase specificity and sensitivity touchdown PCR[162] was used where annealing temperature was gradually decreased by 1^{0}C in sequential cycles.

Components	Volume 1x (µl)
dH_2O	18.8
10x PCR buffer with MgCl2 [15 mM]	2.5
50x dNTP mixture [10 mM each]	0.5
Taq DNA polymerase [5U/µl]	0.2
Forward primer [10 pmol/µl]	0.5
Reverse primer [10 pmol/µl]	0.5
DNA [20 ng/µl]	2

5 min	94^oC	Initial denaturation	
30 sec	94^oC	Denaturation	35 cycles
30 sec	$63\text{-}55^oC^*$	Primer annealing	
1 min	72^oC	Elongation	
5 min	72^oC	Final elongation	

Product verification

For the agarose gel electrophoresis 5 µl of DNA product was mixed with the same volume of 1x bromphenole blue to make the solution visible under natural light for easier tracking during loading and electrophoresis. 6 µl of Ethidium Bromide (EtBr) was incorporated into the 2% agarose gel for band visualization under ultraviolet light.

AMPure cleaning

AMPure Kit was used for the purification of the amplified product from the residual PCR components. The procedure is based on 'Solid Phase Reversible Immobilization' (SPRI) technology[163] which relies on binding of negatively charged DNA molecules to paramagnetic beads. The DNA coupled beads were retained in the reaction wells by placing the reaction plate on a 96-well magnetic plate during the removal of the residual components. The AMPure cleaning was performed according to the following protocol:

1. 36 µl of AMPure solution is added into each 20 µl of PCR mixture and pipette mixed 10 times.
2. The 96-well plate is placed on a magnetic plate and incubated for 10 minutes.
3. The solution is discarded and 200 µl of 70% EtOH is added into each well and discarded after 30 seconds. The ethanol washing is performed twice.
4. The DNA bounded magnetic beads are air-dried for 10 minutes.
5. 40 µl of TE-4 buffer is added into each well for elution and pipette mixed to ensure complete release of DNA molecules from the magnetic beads.

Cycle sequence reaction

Cycle sequencing reactions were performed using the BigDye® Terminator v3.1 Cycle Sequencing Kit. The reaction mixture with a 20 µl of total volume/per reaction and the cycle set up were as below, respectively.

Components	Volume 1x (µl)
dH_2O	15.25
5x Big Dye Terminator v3.1 sequencing buffer	3.75
Big Dye - ready reaction mix v3.1	0.25
Primer F (or R) [3.2 pmol/µl]	0.5
AmPure cleaned DNA template	0.25

1 min	96°C	Initial denaturation	
10 sec	96°C	Denaturation	
5 sec	50°C	Primer annealing	25 cycles
4 min	60°C	Elongation	

CleanSeq cleaning

This step was similar to the first AMPure cleaning and also employed the paramagnetic beads. The CleanSeq cleaning was performed according to the following protocol:

1. 10 µl of CleanSeq solution and 62 µl of 85% EtOH are added into each cycle sequencing mixture and pipette mixed 7 times.
2. The 96-well plate is placed on the magnetic plate and incubated for 3 minutes.
3. The solution is discarded and 100 µl of 85% EtOH is added into each well and discarded after 30 seconds.
4. The DNA bounded magnetic beads are air-dried for 10 minutes.
5. 40 µl of dH_2O is added into each well for elution and incubated for 5 minutes to ensure complete release of DNA molecules from the magnetic beads

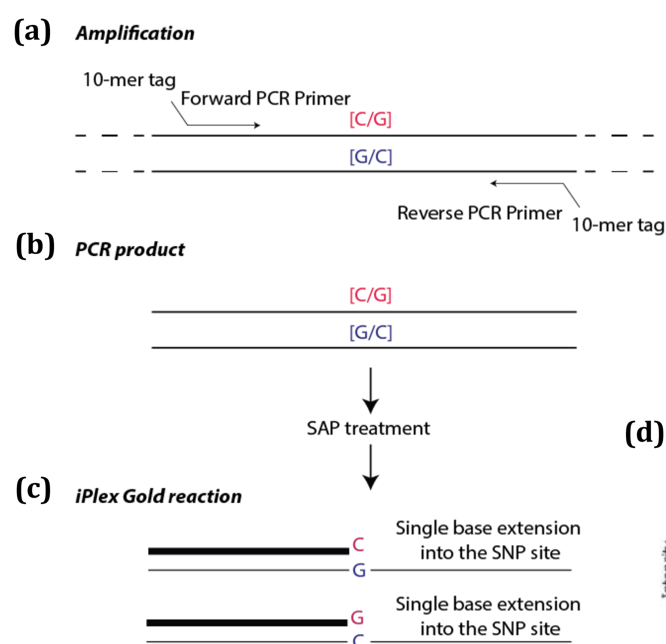
Capillary electrophresis and data assembly

Applied Biosystems 3130xl 16-capillary Genetic Analyzer was the device used for electrophoresis and data generation. The electropherograms were visualized and analyzed by the Seqman II (DNA Star) software.

2.6.5 Genotyping based on MassExtend Reaction (Sequenom®)

Fine mapping or validation of candidate genes necessitates the genotyping of pre-selected groups of variants. MassExtend Reaction combined with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS)[164, 165] by Sequenom is one of the state-of-art technologies for this purpose. This technology works with Sequenom's iPLEX® Gold assay which allows simultaneous analysis of up to 40 variants in a single reaction

(plex). For this application 3 primers are designed per variant. The first two compose a standard primer pair (forward and reverse) for amplification of the targeted region which includes the polymorphic site of interest and flanking sequences (Figure 2.4 a). The third is called an unextended primer [20] and is designed to bind upstream of the polymorphic site with its last base being just adjacent to the polymorphic site (Figure 2.4 c). The initial step in this genotyping method is the amplification of all the targeted regions by PCR in a single plex by simultaneous use of all the standard primers (Figure 2.4 a). After the PCR, deactivation of residual dNTPs is necessary to prevent their incorporation during the single-base-extension (SBE) reaction. This is achieved by shrimp alkaline phosphatase (SAP) mediated dephosphorylation of the dNTPs (Figure 2.4 b). The following step is the SBE reaction where all UEP primers are employed simultaneously to bind upstream of and exactly adjacent to the targeted polymorphic sites. UEPs are then extended at the 3' end by a single ddNTP complementary to the base at the point of variation (Figure 2.4 c). As the SBE reaction is terminated, extension products with allele-specific differences in their masses are obtained. Therefore; when preparing a genotyping assay the primers for tagging a group of variants are so designed that each unextended primer [20] has a unique mass just like each extension product has a unique pre-calculated mass. This way; products with different masses can be detected by mass spectrometry and analyzed for assessment of genotypes (Figure 2.4 d). Resin purification is performed after the SBE reaction to remove the cationic molecules from the assay since the final analysis takes place in an electrical field. The reaction products (analytes) are spotted on a SpectroCHIP which is subjected to MALDI-TOF-MS by the Sequenom MassARRAY™ Compact Analyzer system.



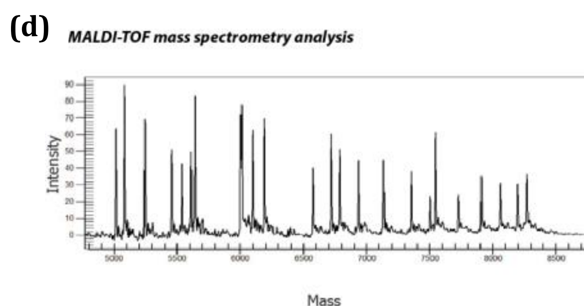


Figure 2.4 Schematic representation of the genotyping protocol based on the MassExtend Reaction (Sequenom®). Amplification of the target region (a) is followed by shrimp alkaline phosphatase treatment for deactivation of residual dNTPs (b). Primers exactly adjacent to the polymorphic sites bind DNA and are extended at the 3' end by a single ddNTP complementary to the base at the point of variation (c). When the single-base-extension (SBE) reaction is terminated, each unextended primer [20] and extension product has a unique mass which can be detected by mass spectrometry (d). Reproduced from the iPLEX® Gold Application Guide (www.sequenom.com)

The MALDI-TOF-MS necessitates the application of a short laser impulse that leads to ionization of the analytes by the assistance of the chip matrix. This leads to their detachment (desorption/ionization) from the matrix. The ionized and matrix-detached molecules are then accelerated in the vacuum flight channel of the mass spectrometer under an electric field[166, 167]. Depending on their individual masses, it takes each analyte a different flight duration to reach the detector; hence the higher the mass of an analyte, the longer is the flight time. The time-of-flight measurements are finally converted to mass information and thereafter to the genotype call information by the system's analysis software.

Primer Design

The primers were designed either by the PreXTEND-Tool in the MySequenom homepage and the Assay Design 3.1 Software or directly by the Assay Design Suite in MySequenom homepage. In the former approach, the standard primer pairs and the UEPs were designed by the PreXTEND-Tool and the Assay Design 3.1 Software, respectively. To serve as input for the PreXTEND-Tool a file was used which consisted of 300 bp long sequences spanning point of variation and its flanking up- and downstream sequences for each variant. Primers were designed to yield amplicons of 80-120 bp in size. The output from the PreXTEND-Tools served as the input file for the Assay Design 3.1 Software for designing the UEP primers. In the latter approach all primers were designed by The Assay Design Suite. The input file was as defined for the PreXTEND-Tool and the amplicons were also designed to yield amplicons of 80-120 bp in size.

Amplification

The genotyping protocol was performed in 384-well plates by using a total of 15 ng dried DNA from each individual. A primer mix was prepared which compromised of 500 nM of each standard primer pair (forward and reverse) that were included in the assay. Beckman NX-MC robot was used for the distribution of the master-mix solution into individual wells (5µl/well). The master-mix solution for a single PCR and the cycle set up were as below, respectively.

Components	**Volume 1x (µl)**
dH_2O	2.7
10 x PCR buffer, MgCl2 [15 mM]	0.625
dNTP mixture [10 mM each]	0.25
MgCl2 [25 mM]	0.325
Standard primer mixture [500 nM each]	1
Hot Star Taq Polymerase [5U/ µl]	0.1

15 min	95°C	Initial denaturation	
20 sec	95°C	Denaturation	45 cycles
30 sec	56 °C	Primer annealing	
1 min	72 °C	Elongation	
5 min	72 °C	Final elongation	

Shrimp alkaline phosphatise digestion

Beckman NX-MC robot was used for the distribution of the SAP reaction mixture into individual wells (2µl/well). The SAP reaction mixture for a single reaction and the cycle set up were as below.

Components	Volume 1x (µl)
dH_2O	1.53
10x SAP buffer	0.17
SAP enzyme [1U/ µl]	0.3

40 min	37°C	SAP digestion
5 min	85°C	SAP inactivation

Single base extension reaction

As a prior step to SBE, an adjustment of the UEP primers was made to assure generation of signals with comparable intensities from low and high mass primers during the mass spectrometry. This adjustment divides primers into mass groups and ensures a higher concentration of high mass primers in the final UEP primer mixture. The primers used in this study were divided into 4 mass groups with the concentration adjustments as given in Attachment IV. Beckman NX-MC robot was used for the distribution of the SBE reaction mixture into individual wells (2µl/well). The SBE reaction mixture for a single reaction and the cycle set up were as below, respectively.

Components	Volume 1x (µl)
dH2O	0.619
10x iPLEX Gold buffer	0.2
iPLEX Termination mix	0.2
UEP primer mixture	0.94
iPLEX Gold enzyme (sequenase)	0.041

30 sec	94°C	Initial denaturation		
5 sec	94°C	Denaturation	1 x	40 cycles
5 sec	52 °C	Primer annealing	5 cycles	
5 sec	80 °C	Primer extension		
3 min	72 °C	Final extension		

After SBE reaction, products were diluted by addition of 16 µl dH_2O into each well and purified by mixing with an ion exchange resin. Resin was spread out on a 384-well dimple plate filling

each hole with about 6 mg of resin. Dimple plate was used to fill in the wells of the 384-well reaction plate. The reaction plate was rotated for 7 min with an overhead shaker to ensure complete purification and then centrifuged for 7 min at 4000 rpm. The reaction products were then spotted on a SpectroCHIP (Sequenom) by the use of the Nanodispenser device. The chip was placed into the Sequenom MassARRAY™ Compact Analyzer system and processed using the RT-workstation software 3.3. The genotype calls automatically generated by the system's analysis software Typer 3.4 / 4.0 were always assessed and confirmed by manual inspection before exporting the final data. A representative genotype analysis is given for a low-frequency variant in the *CHD1L* gene in Figure 2.5.

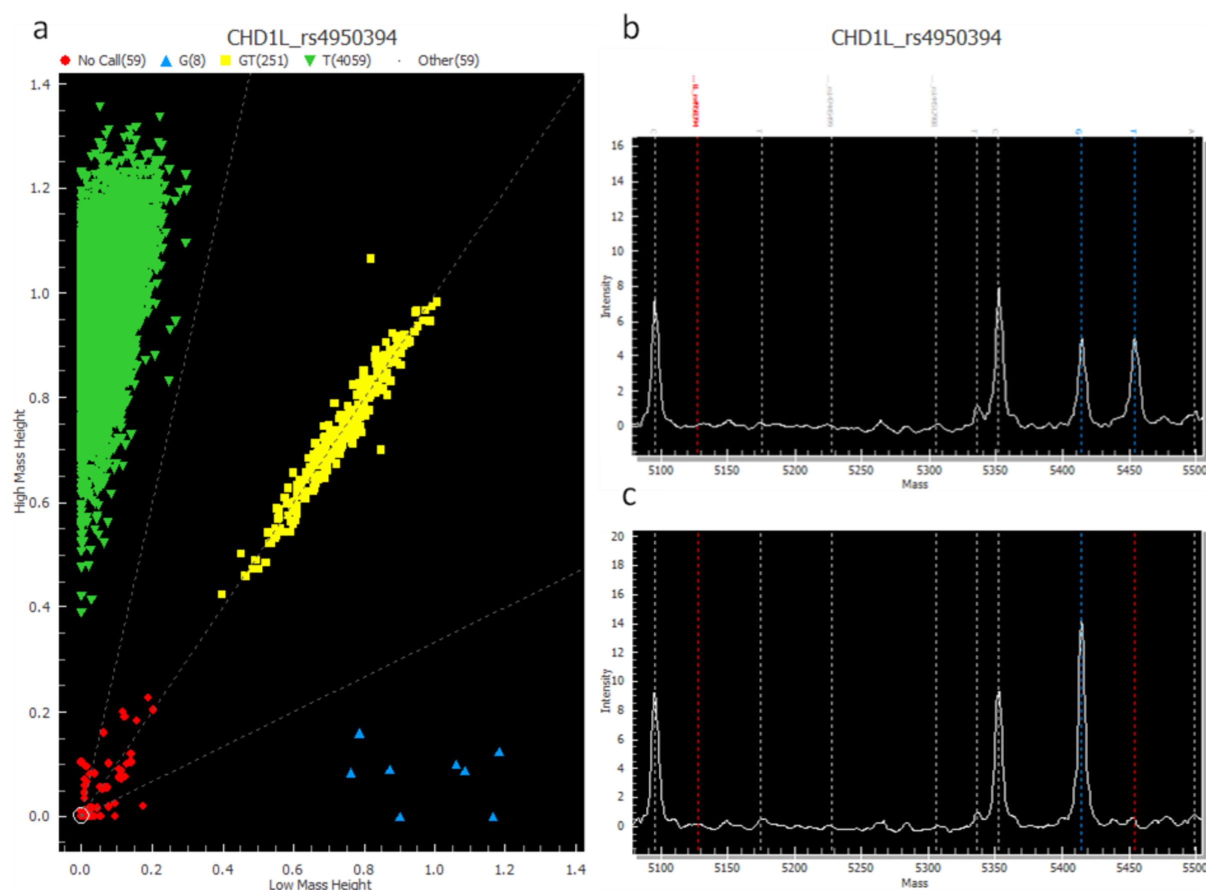


Figure 2.5 A representative genotype analysis with MassExtend technology (a) The cluster plot for rs4950394 in *CHD1L* gene showing individuals carrying the homozygous 'TT' genotype (green triangles), the heterozygous 'GT' genotype (yellow squares) and the homozygous 'GG' genotype (blue triangles). The red circles represent reaction wells where no genotype call was assigned. (b) Mass spectrum representing the masses measured from an individual carrying the heterozygous genotype for rs4950394. Peaks coinciding with the blue dashed lines represent the presence of the two alternative alleles which generated extension products with different masses. The red dashed line marks the specific mass of the UEP and the absence of a peak coinciding with it indicates the complete conversion of all the UEP primers into extension products. The dashed grey lines represent masses of the UEP primers and the extension products from other variants included in the assay. (c) Mass spectrum representing the masses measured from an individual carrying the homozygous 'GG' genotype for rs4950394. Only a single peak coinciding with the specific mass of the 'G' allele containing extension product is observed.

2.6.6 Genome-wide gene expression analysis

Genome-wide gene expression by use of microarray technologies enables simultaneous assessment of the abundance of ten-thousands of transcripts from ten-thousands of genes

spread throughout the entire genome. This approach necessitates intact mRNAs isolated from tissue of interest.

RNA integrity assurance by BioAnalyzer

To guarantee a high quality of the total RNA isolated from the pre-mortem human hippocampal segments, the RNA samples were initially subjected to quality control by BioAnalyzer 2100 measurements with the employment of RNA 6000 Nano LabChip® Kits according to the manufacturer's protocol (http://www.chem.agilent.com). The measurement of RNA integrity is based on principles of gel electrophoresis and simultaneous fluorescence analysis applied to a lab-on-a-chip approach[168, 169]. The LabChip® allows parallel assessment of 12 samples loaded into the sample-wells of the LabChip which contains micro-channels pre-filled with polymer and a fluorescent dye. An RNA 6000 ladder standard containing different RNA fragments with a defined range of sizes is run in parallel to the samples to serve as a reference for data analysis. The electrophoresis enables size based separation of the ribosomal RNA (rRNA) species in the total RNA and detection of the bands corresponding to the predominant 28S and 18S rRNAs which is a standard method to assure that RNA has not been degraded by RNases during isolation or sample handling[170]. The intercalation of the dye into RNA molecules enables them to be detected by laser induced fluorescence. Besides the detection of peaks corresponding to the 28S and 18S rRNAs, an RNA integrity number (RIN) is determined for each sample based on a software algorithm which takes into account the whole electrophoretic trace[169]. The RIN ranges from 1 to 10 where '1' denotes highly degraded RNA and '10' denotes very intact RNA. It was previously claimed that a RIN larger than 5 is necessary for an expression microarray study[171]. The quality assessment by Bioanalyzer 2100 showed intact 28S and 18S ribosomal RNA signals for all used RNA samples as well as a RIN>7.9.

Genome-wide expression analysis by Illumina Bead Technology

All total RNA samples (n=148) were used for a systematic chip-based gene expression analysis by using the Illumina® HT12-v3 Expression BeadChips. Each BeadChip contains 12 microarrays enabling parallel analysis of 12 samples. Illumina® HT12-v3 expression microarrays target more than 25,000 annotated genes with more than 48,000 probes designed by using the RefSeq (Build 36.2, Rel 22) and the UniGene (Build 199) databases. Transcript tagging oligonucleotide probes are covalently attached to beads held in micro-wells on the surface of an array substrate. Hundreds of thousands of copies of a probe is attached to each bead type (Figure 2.6). A high level of bead redundancy (each bead type being represented on average 15 times) improves the data quality and reproducibility. The beads are randomly self-assembled into the micro-wells during manufacturing and owing to the presence of a 29-mer address sequence linked to each

bead; their location can be determined by a hybridization based procedure allowing the mapping of the array.

The workflow for chip-based gene expression analysis is mainly composed of the following stages; preparation and quantification of biotin-labeled cRNA from the total RNA, hybridization of the biotin-labeled cRNA to the BeadChip, immunohistochemical staining by Cy3-conjugated Streptavidin which has high affinity for biotin, laser scanning of the BeadChips leading to fluorescence emission from the Cy3 and determination of the signal intensity values for each bead type (Figure 2.6, top panel). The average signal intensity from each bead type is proportional to the quantity of the respective transcript in the original sample (Figure 2.6, bottom panel).

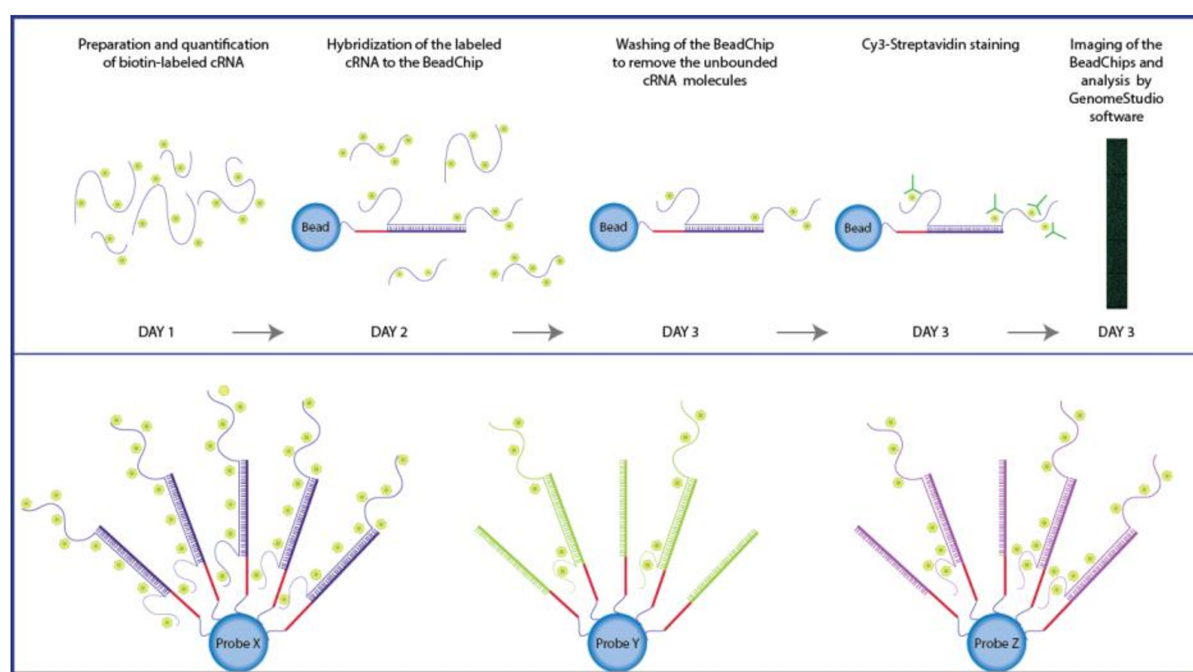


Figure 2.6 Whole-Genome Gene Expression by Direct Hybridization Assay with Illumina BeadArray technology The workflow of the direct hybridization assay (top panel). Schematic representation of three bead types each containing copies of an oliogonucleotide probe tagging a different transcript (bottom panel). For simplicity reasons only several copies of a probe is illustrated to be attached to each bead instead of the hundreds of thousands of copies in reality. The biotin-labeled cRNAs are hybridized to the probes with the amount of hybridization being dependent on the relative abundance of the respective cRNA molecule (bottom panel). Created based on the illustrations from Whole-Genome Gene Expression Direct Hybridization Assay Guide (http://support.illumina.com).

50 ng of total RNA from each sample was converted into biotin-UTP labelled cRNA by using the Illumina TotalPrep-96 RNA Amplification Kit according to manufacturer's protocol (http://tools.lifetechnologies.com/). Labelled cRNA was hybridized to Illumina® HT12-v3 expression beadchips according to the manufacturer's protocol (http://support.illumina.com).

All expression profiles were analyzed using the GenomeStudio software. Quality control and quantile normalization was performed using the R statistical software and packages from the Bioconductor project[172, 173]. The threshold for background signal intensity was calculated as

follows: First a linear regression model between normalized expression values and corresponding detection p-values was fitted. Then the estimated signal intensity at a detection p-value of 0.05 was used as the background threshold to distinguish between signals above and below the background signal. The expression analysis was performed in collaboration with Dr. Andrea Hofmann from the University of Bonn.

2.6.7 Variant detection and bioinformatical analysis

The generated sequence information was screened for any variation of hetero- or homozygous nature by comparing the data with the reference sequence in UCSC Genome Browser based on NCBI build 36.1/hg18 or GRCh37/hg19. For consistency purpose, the given genomic positions of all defined variants in this thesis are based on GRCh37/hg19. Reference SNP ID numbers (rs#) are based on dbSNP build 135 and variants which are not present in dbSNP 135 are designated as novel. RefSeq definition of all targeted transcripts in the 1q21.1 microdeletion region, *NRXN1* gene and *TCF4* gene and their respective proteins in Uniprot are given in Attachment I. In the 1q21.1 microdeletion region, 7 of the 10 genes spanned by the minimally deleted region were targeted and the remaining 3 which are lying in the segmental duplications could not be targeted by sequencing (Attachement II).

In the 1q21.1 microdeletion region variants with a minor allele frequency (MAF) of <5% in the combined patient and control sample were defined as low-frequency variants whereas in the *NRXN1* study, variants with a MAF<3% in the combined sample were considered. In the *TCF4* study, only schizophrenia patients were resequenced and the sequence information from 379 European individuals from 1000 Genomes Project was used to account for rare variation in the control population[154]. The sequence variants observed in the 379 European individuals from the 1000 Genomes Project were retrieved from the 1000 Genomes Browser by screening of the VCF files for the regions targeted and analyzed in the resequencing step (coding exons and 50 bp up- and down-stream flanking sequences of the *TCF4* transcript variant 1, NM_001083962.1). Variants with a MAF<3% either in the patient sample or in the 379 European individuals of 1000 Genomes Project were defined as low-frequency variants.

All identified variants were primarily subjected to bioinformatical analysis by use of two main *in silico* tools: Mutation Taster; to assess whether they were leading to an amino acid substitution (missense variant) or a premature stop codon (non-sense variant) and Human Splicing Finder (HSF); to assess whether they were leading to an acceptor/donor site change (splice site variant). Expasy Translate tool was used for codon translation for assessment of the consequences of the splice site changes. Missense variants were further evaluated by PolyPhen-2 and SIFT to assess their potential impact on protein function. Localization of missense variants in protein domains in 1q21.1 microdeletion region and NRXN1 were defined according

to Uniprot, Expasy and SMART tools. The locations of missense variants in TCF4 protein domains were defined based on a recent study by Sepp *et al.* where comprehensive structural and functional characterization of TCF4 isoforms was reported[174]. Secondary structure of all proteins was assessed by the Jpred3 and Dompred tools. All variants were assessed for evolutionary conservation at nucleotide level among human, chimp, rhesus, monkey, and dog and for localization within transcription factor binding sites by the 'Vertebrate Multiz Alignment and Conservation' and the 'HMR Conserved Transcription Factor Binding Sites' tracks of the UCSC Genome Browser, respectively.

2.6.8 Statistical analysis

NRXN1 gene

For the analysis of the sequence data from *NRXN1* gene, a two-tailed Fisher's Exact test was applied which was one of the state-of-the-art approaches by the time of this study. Fisher's Exact test was applied to evaluate the overall burden of low-frequency sequence variants in patients and controls. The analysis was performed by collapsing the low-frequency variant information on individual level to compare the total number of patients carrying a low-frequency variant with the total number of controls carrying a low-frequency variant. The analysis was applied on i) the entire variant set and ii) subsets of variants stratified by functionality (i.e. potentially functional -missense, splice site, non-sense- vs. nonfunctional -intronic and synonymous- variants). The genotyping data was also analyzed by the Fisher's Exact test for comparing the burden of individual variants or all of the genotyped variants in patients and controls.

1q21.1 microdeletion region

Several statistical methods developed for rare variant association testing in a genomic region were employed for the analysis of the data from initial sequencing sample and genotyping-1 sample of the 1q21.1 microdeletion region. The rare variant association tests can mainly be divided into burden and non-burden tests. Burden tests collapse the information of rare variants in a genomic region into groups and then test the association of the phenotype with these collapsed groups of variants. Although the main logic in burden tests is similar, these tests differ from each other by incorporating other types of information or accounting for other criteria. Three burden tests were applied:

-the Combined Multivariate and Collapsing approach by Li and Leal[175] which collapses SNPs into subgroups and subsequently applies multivariate analysis;

-the weighted-sum statistic test by Madsen and Browning[176] which performs a rank-sum test on the individual weighted genotype scores;

-the burden test developed by Price *et al.*[177] which can utilize external information on the rare variants such as their respective PolyPhen scores.

More recently, several statistical tests differing from traditional burden tests by accounting for opposite effect directions (protective or risk) of variants have been proposed. From these classes of tests the following were applied:

- the replication based strategy of Ioanita-Laza *et al.*[178] which also pools variants together, however conditioned on the different sharing patterns of rare alleles between cases and controls;
- SKAT by Wu *et al.*[179] which is based on a kernel machine regression framework and assesses similarities between genotype patterns of individuals across multiple markers in a region of interest and tests whether there is a relation between genotypic similarities and phenotypic similarities;
- the C-Alpha test by Neale *et al.*[180] which analyzes the distribution of rare variants observed in cases versus controls;
- the distance-based measure (DBM) method developed by Fier at al.[181], which analyzes the differences in the spatial clustering of rare variants observed in cases and in controls.

The described methods were applied on regional genotype data prepared as a matrix of allele counts across the region of interest (0=non-carrier, 1=heterozygous carrier, 2=homozygous carrier of the minor allele). In the regional analysis, individuals who yielded incomplete genotype matrices by failing a genotype call for at least one variant were excluded from the analysis. In the initial sequencing sample, 2 patients were excluded from the analysis based on this criterion, leaving a total of 92 patients and 94 controls which were included in the analysis of the resequencing data. In the genotyping-1 sample, 4 patients and 3 controls were excluded based on the same criterion leaving a total of 1896 patients and 2183 controls which were included in the analysis of the genotyping data. The analyses were performed at 1, 3 and 5% MAF cut-offs and applied on i) the total sample (resequencing and genotyping samples combined) and ii) only the genotyping sample. Significance was assessed empirically by using permutation. Briefly, the test statistics were first calculated for the original dataset, then the case-control status in the sample was randomly permutated 1000 times to re-calculate the test statistics for every permutated sample and finally empirical p-values were derived based on the generated distribution of the test statistics. The case/control ratio in the original dataset was kept constant across permutations. Whenever significant associations were detected ($p<0.05$); the analyses were re-ran with 10,000 permutations to confirm the signal. Single marker analyses were performed by Pearson's Chi-square test to detect significant differences in the allele distributions of single markers for cases and controls. Significance was assessed empirically by application of 1000 permutations of the case-control status as described.

Whenever a significant p-value below 0.05 was detected the analysis was repeated by 10,000 permutations.

The association testing of prioritized variants (genotyping-2 sample) was performed by Pearson's Chi-square test as described and also by logistic regression controlling for ancestry as the genotyping-2 sample was composed of different populations (i.e. German, Danish and Dutch). The empirical p-values for logistic regression analysis were derived from 1000 permutations where the specific case/control ratios of every population were kept constant.

All the analyses were performed in R software and in collaboration with Dr. Heide Fier from the University of Bonn.

TCF4 gene

The regional genotype data from *TCF4* locus was analyzed by application of the same rare variant association tests described for the 1q21.1 microdeletion region. In the resequencing sample which constituted of 190 schizophrenia patients, all individuals yielded complete genotype matrices. In the genotyping sample 1 patient and 3 controls were excluded from the analysis due to incomplete genotype matrices leaving a total of 1807 patients and 2258 controls which were included in the analysis. The analyses were performed at 0.3, 1 and 3% MAF cut-offs and applied on i) the total sample (resequencing and genotyping samples combined) and ii) only the genotyping sample. In the total sample analysis the genotypes of the 379 European individuals from the 1000 Genomes database were not included to avoid any bias from population specific rare variant patterns.

3. RESULTS

3.1 Resequencing of the *NRXN1* gene

By sequencing the coding exons and flanking sequences of the *NRXN1* gene in 94 schizophrenia patients and 94 healthy controls about 10.9 kb genomic sequence per individual was generated and analyzed, totalling 2.05 Mb of sequence information in the whole sample. The resequencing analysis of *NRXN1* revealed a total of 21 variants; all of which were single base exchanges. Five of the identified variants were denoted to be common (MAF≥3%) and all were previously annotated in the dbSNP database. The remaining 16 variants were denoted to be low frequency (MAF<3%) variants. Four of the low-frequency variants were leading to amino acid substitutions and they were classified as 'potentially functional variants' whereas the other 12 were synonymous or intronic variants with no direct effect on primary protein structure (Table 3.1).

Table 3.1 Low-frequency variants observed in *NRXN1* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes[a]	Effect (SIFT/PolyPhen)	Conservation[b]	MAF (≤)[c]	Minor allele counts (n=patients/controls)
51254810	rs112934082	T>C	exon 2	E210G	tolerated/benign	5	0.01	1/0
51254577	rs111501521	G>A	intron 2	-	-	2	0.01	1/0
51153131	rs112297733	A>G	intron 3	-	-	5	0.01	1/0
50858257	rs113989332	T>G	intron 5	-	-	5	0.01	1/0
50850686	rs2303298	G>A	exon 7	syn	-	5	0.01	1/2
50848407	rs111648327	G>A	intron 7	-	-	5	0.01	2/1
50847195	rs78540316	G>A	exon 9	P469S	tolerated/probably damaging	5	0.01	1/1
50847150	rs113028018	C>T	intron 9	-	-	3	0.01	1/0
50765412	rs56086732	G>T	exon 11	L748I	tolerated/benign	5	0.01	0/2
50758613	rs111940222	G>A	intron 11	-	-	5	0.01	1/0
50699598	rs112638127	T>C	exon 17	I1068V	tolerated/benign	5	0.01	1/0
50574010	rs113067443	G>T	exon 1 (β)	syn	-	1	0.01	1/0
50573817	rs13023114	T>C	intron 18	-	-	4	0.03	1/3
50281894	rs74746635	A>G	intron 21	-	-	4	0.01	3/0
50149214	rs112536447	A>G	exon 24	syn	-	4	0.01	0/1
50149133	rs113380721	C>T	exon 24	syn	-	4	0.01	2/1

[a]Amino acid positions refer to the NRXN1 α2-isoform with 1547 residues except for G26G which refers to the β-isoform encoding 442 residues.
[b]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[c]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency. The table is modified from Mühleisen *et al.,* 2011[150].

Two of the missense variants (E201G and I1068V) were detected as singleton observations in patients and both were annotated to be benign by SIFT and PolyPhen tools. P469S was predicted to be tolerated and probably damaging by SIFT and PolyPhen, respectively, and was observed in a patient and a control individual. L748I was observed in two control individuals and was predicted to be benign by both tools. The localization of the missense variants on the NRXN1 protein domain structure is given in Figure 3.1. From the remaining 12 variants six intronic and one synonymous variant were observed only in patients, one synonymous variant was observed in a single control individual and four were shared by both groups.

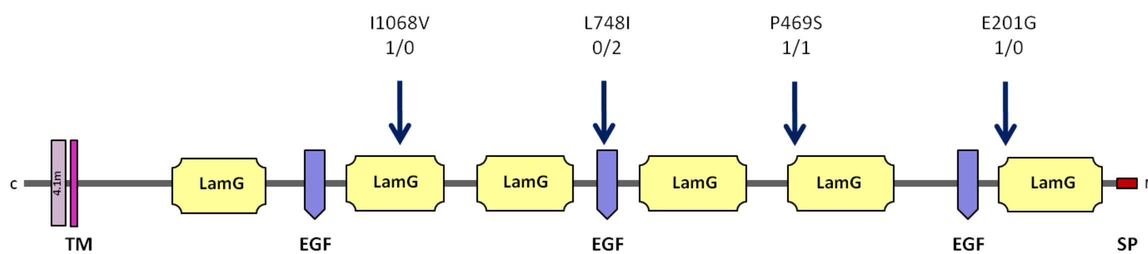


Figure 3.1 Protein domain structure of NRXN1 α2-isoform and localization of observed rare missense variants. Numbers of observations in the total sample are given for patients and controls ($n_{patients}/n_{controls}$). TM, transmembrane domain; LamG, laminin G domain; EGF, epidermal growth factor-like domain; SP, signal peptide. The figure is modified from Mühleisen *et al.,* 2011[150].

Fisher's exact test was used to evaluate overall burden of low-frequency variants in patients and controls. When considered all together, the low-frequency variants were insignificantly overrepresented in patients compared to controls (18 patients/11 controls; *P*=0.23). The overrepresentation mainly arose from variants of no direct effect on primary protein structure (15 patients/8 controls; *P*=0.18) and the missense variants were equally observed in patients and controls in total (3 patients/3 controls; P=1.0).

The potentially functional –missense- variants (E201G and I1068V) which exclusively were observed in patients were considered to be the most plausible candidates for moderate to high risk effects. Both of these substitutions were at highly conserved regions and both were located in different Laminin G domains (Figure 3.1).

E201G was a non-conservative substitution of a polar glutamate to a neutral, non-polar glycine residue. The patient who carried the E201G substitution had an age of onset of seven years and she had a brother with a history of depression. Parental DNA was available only from the mother who did not carry the variant. No information was available concerning possible psychiatric disease in the father or paternal relatives. I1068V was a conservative substitution with no alterations in polarity or charge. I1068V was located in a beta sheet structure and the substitution did not lead to a change in the secondary structure. The patient who carried the I1068V had an age of onset of 20 years. Sequencing of parental DNA revealed that the variant was inherited from the unaffected father whose second degree relatives were also unaffected. These two missense variants were followed up by genotyping in an additional 1415 patients

and 1167 controls to investigate whether any of them were true individual risk variants. The genotyping revealed a control individual who carried E201G (0 patients/1 control, P=0.45). There were no additional individuals who carried I1068V. No evidence for overrepresentation of the potentially functional variants in patients or in controls was found when the discovery sample was also included in the analysis, neither when all the variants were considered together (2 patients/1 control, P=1) nor when they were analyzed individually (E201G: 1 patient/1 control, P=1; I1068V: 1 patient/0 controls, P=1).

3.2 Resequencing of the 1q21.1 microdeletion region

By sequencing the coding exons and flanking sequences of the 7 genes in the 1q21.1 microdeletion region in 92 schizophrenia patients and 94 healthy controls about 28.4 kb genomic sequence per individual was generated and analyzed, totalling 5.3 Mb of sequence information in the whole sample. Altogether, 87 sequence variants were identified in the whole region encompassing the exons and flanking sequences of the seven targeted genes. The majority of the detected DNA changes (n=79) were single base exchanges and the rest were indels (n_{del} = 6, n_{ins}=2) ranging in size from 1 to 4 basepairs. All of the indels were intronic with no potential functional consequences. Of the detected sequence variants, 32 were defined as common with a MAF≥5%. All of the common variants were previously annotated in dbSNP database. A total of 55 variants were defined as low-frequency variants with MAF<5%. Among the low-frequency variants, 46 had a MAF<3% and 34 had a MAF<1%. Of the 55 low-frequency variants, 32 were synonymous or intronic with no direct effect on primary protein sequence, one was an exonic variant predicted to create a new splice site by the Human Splicing Finder tool, 21 were missense variants leading to an amino acid substitution, and one was a non-sense variant leading to a premature stop codon. The number of variants (MAF<5%) observed in each gene and the potentially functional ones among them (missense, non-sense) are denoted in Table 3.2.

Table 3.2 The gene-based distribution of low-frequency variants observed in 1q21.1 microdeletion region

	PRKAB2	***FMO5***	***CHD1L***	***BCL9***	***ACP6***	***GJA5***	***GJA8***
Length (AA)	272	533	897	1426	428	358	433
Variants	3	5	21	12	9	1	4
Functional variants	1	3	6*	5	4	0	3

*A variant which was predicted to lead to a splice site change by HSF was observed in *CHD1L* however it was not included in the functional variants counts before validation of the *in silico* predicted effect. AA, amino acid.

PRKAB2

In the *PRKAB2* gene, three variants were observed, one was a missense variant and two were synonymous or intronic variants (Table 3.3). The missense variant was observed in a single patient and was predicted to be damaging by both SIFT and PolyPhen. It was in a highly conserved region and was located at a transcription factor binding site for GATA-1. The substituted amino acid was located in a beta sheet structure and the substitution was not predicted to lead to a change in the secondary structure by *in silico* analysis. The synonymous and intronic variants were observed both in patients and controls and were in complete LD in the present sample.

Table 3.3 Low-frequency variants observed in *PRKAB2* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)
146643580	rs34838459	G>A	exon 2	syn	-	3	0.05	7/8
146643495	rs72708505	C>A	intron 2	-	-	4	0.05	7/8
146639424	-	C>T	exon 3	R82H	damaging/probably damaging	5	0.01	1/0

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency.

FMO5

In the *FMO5* gene, five variants were observed. One of these was a non-sense variant, two were missense and two were intronic variants (Table 3.4). The non-sense variant (R485*) was detected in a single patient. It was located in a highly conserved region and led to a premature termination codon (PTC) in exon 9 of *FMO5* which may result in production of a truncated protein. Non-sense mediated mRNA decay (NMD) might be avoided due to the location of the PCT in the terminal exon[182-185]. The predicted truncated protein lacks ~9.2% of the amino acid content which includes a hydrophobic segment at the C-terminal (PRINTs database accession number: PR01125, Motif IX) which had previously been suggested to be responsible from membrane anchorage of the enzyme[186].

Table 3.4 Low-frequency variants observed in *FMO5* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)
146684095	rs58351438	T>C	exon 5	K166E	tolerated/probably damaging	5	0.03	3/1
146680394	rs6684454	C>T	intron 6	-	-	5	0.05	7/8
146672745	rs56134376	T>C	exon 7	Q391R	tolerated/benign	5	0.01	0/2
146661723	rs184393893	G>A	intron 8	-	-	4	0.01	1/1
146658628	-	G>A	exon 9	R485*	n.a./n.a.	5	0.01	1/0

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; n.a., not available; MAF, minor allele frequency.

One of the missense variants predicted to be benign by both PolyPhen and SIFT (Q391R) was located at an alpha-helix and was observed only in two controls. The other missense variant (K166E) was predicted to be probably damaging only by PolyPhen and was observed in three patients and a control individual. The substituted amino acid was the first residue of a sequence motif which was suggested to contribute to the NADPH binding of FMO5 and was expected to have an additional function as well[187, 188]. The intronic variants observed in *FMO5* were carried both by patients and controls.

CHD1L

In the *CHD1L* gene, which is the second largest gene in the region, the highest number of variants (n=21) was observed (Table 3.5).

Table 3.5 Low-frequency variants observed in *CHD1L* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)
146736063	-	T>C	intron 6	-	-	4	0.01	0/1
146736068	-	TTAT>-	intron 6	-	-	5-5-3-5	0.01	0/1
146736137	rs36008075	A>G	exon 7	-	-	4	0.03	4/1
146737517	-	->T	intron 7	-	-	3-3[i]	0.01	1/0
146737540	rs140555192	C>T	intron 7	-	-	4	0.05	5/7
146740514	-	G>A	exon 10	syn	-	4	0.01	1/1
146742648	rs144757186	G>A	exon 11	D381N	damaging/probably damaging	5	0.01	1/0
146747069	rs2275250	T>C	exon 13	syn	-	4	0.01	1/0
146747965	rs185219867	C>T	Intron 14	-	-	4	0.03	3/2
146751782*	rs7547279	C>A	exon 15	syn	-	1	0.03	2/7
146756234	-	CAA>-	intron 16	-	-	3-2-3	0.01	1/0
146757032	-	C>A	exon 17	T629N	tolerated/benign	4	0.01	0/1
146757132	rs142236750	G>A	exon 17	syn	-	5	0.01	2/0
146757200	-	G>A	intron 17	-	-	4	0.01	1/0
146758054	rs139791996	G>A	exon 17	G700R	tolerated/benign	3	0.01	1/0
146759387	rs148289715	A>G	exon 19	I765M	tolerated/benign	5	0.01	1/1
146759428	rs144512908	G>A	intron 19	-	-	5	0.03	1/3
146765379	rs148434783	A>G	exon 21	I827V	tolerated/benign	5	0.01	1/0
146766070*	rs67589628	AT>-	intron 21	-	-	4-1[i]	0.03	2/7
146766122	rs45563244	C>G	exon 22	syn	-	4	0.03	5/3
146767149*	rs4950394	T>G	exon 23	S885A	tolerated/benign	1	0.03	2/7

*The three variants are in complete LD and the low-frequency alleles for them are given in the reference sequence in human genome build 19. The high frequency alleles are given as the substituted residues and the low-frequency alleles are given as the substituting residues in the table for consistency with other variants. The same annotation applies to the amino acid changes. (e.g. For rs4950394, 'G' is the low-frequency allele and amino acid 'A' is encoded in the presence of G allele. The minor allele counts thus refer to the 'G' allele). [a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation). [i]For the insertions the conservation scores of the residues preceeding and postceeding the insertion are given. *rs67589628 is treated as an insertion in the 'Conservation' column although it is annotated as a deletion in the 'Alteration' column. This is because the reference sequence does not contain the 'AT' residues in the USCS Genome Browser (as described with *) and conservation can only be assessed by the preceeding and postceeding bases. [b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency.

Six of the *CHD1L* variants were missense variants, one was predicted to lead to a splice site change by the HSF tool and 14 were synonymous or intronic variants. Of the six missense variants, three were observed only in patients (D381N, G700R, and I827V) as singleton observations. One of these missense variants was located in a highly conserved region (D381N) and was predicted to be damaging by both SIFT and PolyPhen. The substitution was located at an alpha-helix structure and affected the helicase superfamily c-terminal domain of the CHD1L protein (Figure 3.2). The second patient-specific nonsynonymous variant (G700R) was located in a low complexity region and it was the only non-synonymous rare variant (MAF<1%) with some evidence for segregation. It was carried by the index patient and a brother who was also affected by schizophrenia as well as their father who suffered from an antisocial personality disorder. The variant was predicted to be benign both by SIFT and PolyPhen. The third patient specific non-synonymous variant (I827V) was in a highly conserved region and was located at a transcription factor binding site for MEF-2A. The substitution was at an alpha-helix structure at the macro domain of CHD1L and was predicted to be benign by SIFT and PolyPhen.

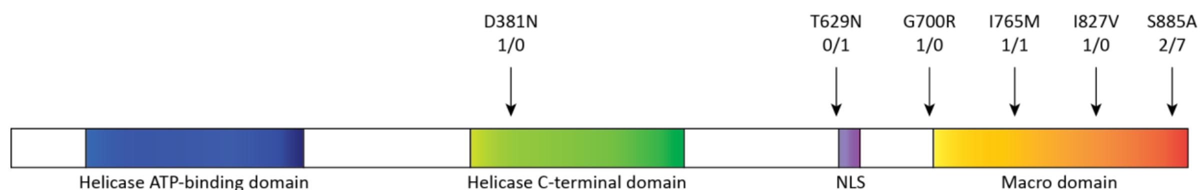


Figure 3.2 Protein domain structure of CHD1L and localization of observed rare missense variants. The protein domain structure is given for CHD1L isoform 1 composed of 897 amino acid residues. Numbers of observations in the total sample are given for patients and controls ($n_{patients}/n_{controls}$). NLS, nuclear localization signal. The domain information retrieved from http://prosite.expasy.org/

Of the remaining three missense variants, one was observed only in a single control individual (T629N) and two which were located at the macro domain were shared by patients and controls (I765M and S885A) (Figure 3.2). All three were predicted to be 'benign' by PolyPhen. Analysis by the HSF tool revealed that a synonymous A to G substitution in exon 7 of *CHD1L* (rs36008075) created a new acceptor site that had a higher consensus value (85.5) than the wild-type site (73.9)[11]. The new splice site lead to an in-frame deletion of 19 amino acids (amino acid residues 193-211) located at the helicase ATP-binding domain. From the remaining 14 synonymous or intronic variants observed in *CHD1L*, five and two were detected only in patients and in controls, respectively and the others were shared between two groups.

BCL9

In the *BCL9* gene which is the largest gene in the region, 12 variants were observed of which 5 were missense and 7 were synonymous or intronic variants (Table 3.6). Two missense variants (P113S and N147H) were observed only in patients as singleton events. P113S was predicted to be benign by both SIFT and PolyPhen whereas N147H was predicted to be damaging by SIFT.

Table 3.6 Low-frequency variants observed in *BCL9* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)
147084965	rs41295833	C>T	exon 5	P113S	tolerated/benign	5	0.01	1/0
147086294	-	A>C	exon 6	N147H	damaging/benign	5	0.01	1/0
147090614	-	G>T	intron 7	-	-	4	0.01	0/1
147090775	rs143556015	C>T	exon 8	R272C	tolerated/benign	5	0.01	0/1
147090956	rs61751616	C>T	exon 8	P332L	tolerated/benign	5	0.01	1/2
147091005	rs61729410	C>T	exon 8	syn	-	4	0.01	2/0
147091416	rs80312516	G>A	exon 8	syn	-	5	0.03	1/5
147091689	rs61751617	G>A	exon 8	syn	-	3	0.05	8/9
147092112	rs61754125	G>A	exon 8	syn	-	5	0.05	9/8
147095577	-	T>C	intron 9	-	-	5	0.01	1/0
147096112	rs61751618	G>A	exon 10	M1211I	tolerated/benign	5	0.03	5/3
147096505	rs77650336	C>T	exon 10	syn	-	4	0.05	7/5

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency.

One missense variant located at a low complexity region and predicted to be benign was observed only in a single control individual (R272C). Two missense variants (P332L and M1211I) were detected both in patients and controls and were predicted to be benign by both Polyphen and SIFT. Among the 7 synonymous or intronic variants, 2 and 1 were detected only in patients and in controls, respectively and the remaining were shared between two groups.

ACP6

In the *ACP6* gene 9 variants were observed of which 4 were missense and 5 were synonymous or intronic variants (Table 3.7).

Table 3.7 Low-frequency variants observed in *ACP6* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)
147131663	rs140662730	AAG>-	intron 2	-	-	5	0.01	1/2
147131553	rs143920833	T>C	exon 3	D146G	tolerated/possibly damaging	5	0.01	2/0
147131103	rs11800736	C>T	exon 4	syn	-	5	0.05	11/6
147131025	rs41295837	G>A	intron 4	-	-	5	0.03	1/5
147126437	rs140566115	G>A	exon 6	R218W	tolerated/probably damaging	5	0.03	4/4
147126416	-	G>C	exon 6	Q225E	tolerated/probably damaging	5	0.01	0/1
147121996	-	C>T	exon 8	syn	-	4	0.01	1/0
147119397	-	C>T	intron 9	-	-	4	0.01	1/1
147119257	rs137987097	T>A	exon 10	T419S	tolerated/benign	4	0.01	2/1

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency.

The only non-synonymous patient specific variant carried by more than one patient was identified in the *ACP6* gene. It was a missense variant (D146G) observed in 2 patients and predicted to be possibly damaging by PolyPhen. Two of the remaining missense variants predicted to be probably damaging (R218W) by PolyPhen only and benign (T419S) by both SIFT and PolyPhen were carried by both patients and controls. Finally a missense variant located at an alpha helix structure and predicted to be probably damaging (Q225E) by PolyPhen was observed in a single control individual. The majority of the synonymous or intronic variants were observed in both groups and only 1 synonymous variant was observed in a single patient.

GJA5 and *GJA8*

In the *GJA5* gene only a single synonymous variant was discovered which was carried by 3 control individuals (Table 3.8). In the *GJA8* gene 3 missense variants and 1 synonymous variant were discovered (Table 3.8). Two of the missense variants (V129I, I247M) were observed in single patients each, the former was predicted to be benign by both tools and the latter was predicted to be possibly damaging by PolyPhen. The third missense variant (N220D) was shared by 2 control individuals and was predicted to be damaging by both of the tools. Mutations in *GJA8* gene are known to be causative of zonular pulverulent cataract (CZP1; OMIM: 116200) and cataract-microcornea syndrome (CAMIS; OMIM:116150]. One of the missense variants (I247V) identified in a patient in our sample was previously defined in a Russian mother and her son with zonular pulverulent cataract[189]. This mutation was suggested to be causative as it was absent in the unaffected family members. However, no phenotypic information from the schizophrenia patient who carried this variant in our sample was available to link the variant to a cataract phenotype.

Table 3.8 Low-frequency variants observed in *GJA5* and *GJA8* genes

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]	MAF (≤)[b]	Minor allele counts (n=patients/controls)	Gene
147230978	rs2232191	G>A	exon 2	syn	-	4	0.01	0/3	*GJA5*
147380467	rs142415337	G>A	exon 2	V129I	tolerated/benign	3	0.01	1/0	*GJA8*
147380740	rs138140155	A>G	exon 2	N220D	damaging/probably damaging	5	0.01	0/2	
147380823	rs80358202	T>G	exon 2	I247M	tolerated/possibly damaging	4	0.01	1/0	
147380886	rs3766503	C>T	exon 2	syn	-	3	0.05	6/8	

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined frequencies in cases and controls in the total sample. AA, amino acid; syn, synonymous; MAF, minor allele frequency.

1q21.1 microdeletion carriers

The targeted genes were also resequenced in two 1q21.1 microdeletion carriers in order to investigate the presence of recessive risk alleles on the non-deleted strand[190]; however no rare variants were identified to be carried by either of the two patients.

Mendelian inheritance

All truly rare potentially functional variants (missense or non-sense) in the 1q21.1 microdeletion region with a MAF<1% detected in patients were tested for Mendelian inheritance when parental DNA was available. All variants were inherited with ~70% of them being transmitted from the father (P=0,087). Only three of the affected patients had a positive family history of schizophrenia and each of them was carrying a missense variant in the *CHD1L* gene. In two of these cases variants were inherited from unaffected parents and in one of the cases a partial segregation was identified as already described above.

3.2.1 Microarray gene expression analysis of genes at the 1q21.1 microdeletion region in human hippocampus

In order to assess their biological plausibility as candidate schizophrenia genes, expression data of the 7 genes were retrieved from microarray based whole transcription analyses of pre-mortem human hippocampus tissue samples. Whole transcription analyses with Illumina® HT12-v3 Expression BeadChips revealed that *PRKAB2*, *CHD1L*, *BCL9* and *ACP6* were definitely expressed in the human hippocampus (Fig. 3.3). *BCL9* and *PRKAB2* had the highest expression levels which were followed by *CHD1L* and *ACP6* with relatively moderate expression levels. The signal intensities from remaining transcripts (*FMO5*=90.05, *GJA5* transcript variant A and B= 87.85 and 90.86, respectively, *GJA8*=85.19) were below the background signal intensity (96.94). *BCL9* showed the highest expression with a mean signal intensity of 170.7 (min = 125.5 and max = 222.9) which was followed by *PRKAB2* with a mean signal intensity of 141.1 (min = 113.3 and max = 253.5).

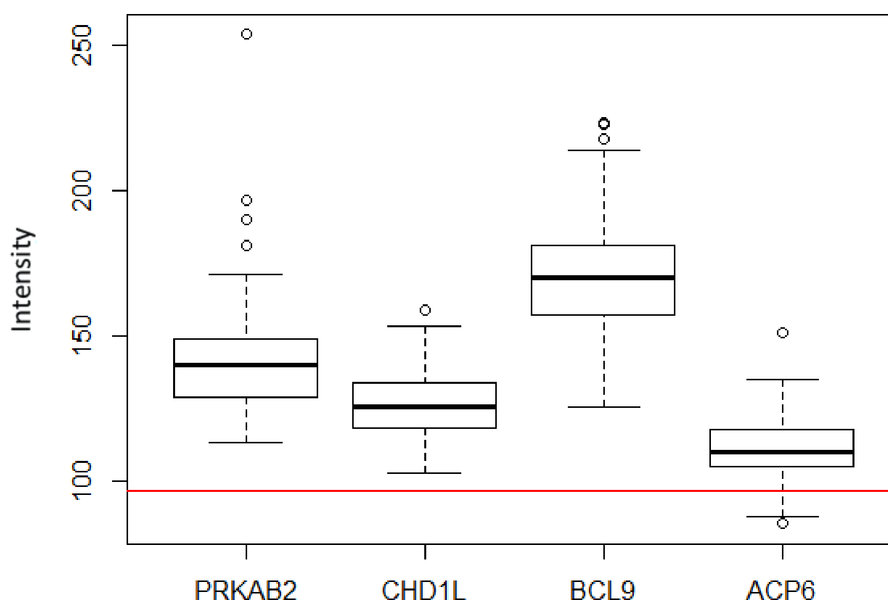


Figure 3.3 Microarray gene expression analyses of genes located in the 1q21.1 microdeletion region. Transcriptional expression analyses of *PRKAB2*, *BCL9*, *CHD1L* and *ACP6* in human hippocampus tissue using Illumina® HT12-v3 expression beadchips (n = 148). The remaining genes are not depicted in the figure as their signal intensities fell below the background signal intensity (96.49) represented with the red line.

CHD1L and *ACP6* were moderately expressed in the human hippocampus with mean signal intensities of 125.8 (min = 102.56 and max = 158.9) and 111.1 (min = 85.37 and max = 151.3), respectively. This data was also compared with the human brain hippocampus microarray data from the Allen Brain Atlas database. The genes identified to be definitely expressed in pre-mortem hippocampus were also shown to be expressed in the Allen Brain Atlas database correlating with the relative expression levels identified in the present dataset (i.e. *BCL9*>*PRKAB2*>*CHD1L*>*ACP6*). Among the remaining genes for which expression could not be confirmed in the present dataset, *GJA8* and *FMO5* were shown to be expressed in human hippocampus with *GJA8* having higher expression levels than *FMO5*. Indeed both *GJA8* and *FMO5* had higher hippocampus expression levels than *ACP6* in the Allen Brain Atlas database. *GJA5* gene yielded signal intensities far below the other genes in this database indicating it is either expressed at very low levels or not expressed at all in human hippocampus which correlates with our findings. Finally, the microarray data from the BrainSpan Atlas was also investigated to look for the expression of these genes in the developmental stages of human brain. This dataset showed the prominent hippocampal expression of *BCL9*, *PRKAB2*, *CHD1L* and *GJA8* genes during the prenatal stages of brain development while the expression levels of *ACP6*, *FMO5* and *GJA5* were shown to be lower.

3.2.2 Association analysis of the resequencing data from 1q21.1 microdeletion region

In order to test association of low-frequency variants in the 1q21.1 microdeletion region several statistical tests developed for association testing of low-frequency variants were applied on the dataset. 1000 random permutations on the case-control status were performed by keeping the number of cases and controls constant to re-generate the test statistics and to derive empirical p-values based on the distribution of these test statistics. Whenever significant associations were detected ($p<0.05$); the analyses were re-ran with 10,000 permutations to confirm the signal. In the following, the method of Li and Leal is referred to as CMC[175], the method of Madsen and Browning as WSS[176], the method of Price *et al.* as RANK[177], the method of Fier *et al.* as DBM[181], the method of Ionita-Laza *et al.* as REP[178], the method of Wu *et al.* as SKAT[179] and the method of Neale *et al.* as C-ALPHA[180]. The initial association analysis; referred to as 'whole region analysis', was performed on the complete low-frequency variant dataset (n=55, MAF≤5%) based on the sequence information from seven genes. The empirical p-values were generated after application of the outlined rare variant association methods at different MAF cut-offs (1, 3 and 5%) as depicted in Table 3.9. The whole region analysis revealed a significant association of low-frequency variants with schizophrenia.

Table 3.9 Association analysis of low-frequency variants in the 1q21.1 region with schizophrenia

MAF	CMC	WSS	RANK	DBM*	REP*	SKAT	C-alpha
0.01	0.616	0.396	0.138	0.086	0.104	0.361	1.000
0.03	0.883	0.717	0.216	**0.021**	0.279	0.131	0.154
0.05	1.000	0.824	0.226	**0.039**	0.496	0.333	0.834

The reported p-values for all test statistics are given for calculations based on 10000 permutations. Significant p-values below 5% significance level are given in bold. The better performing two- and one-sided test statistics for DBM and REP methods are given, respectively.

The association signal was reported by the DBM method both at 3% (*P*=0.021) and 5% (*P*=0.039) MAF cut-offs (Table 3.9). An association below a significance level of 10% was also detected for the 1% MAF cut-off (*P*=0.086) by the DBM.

In order to refine the association signal and identify the signal-driving gene/s in the region, the analysis was performed on single gene level (Table 3.10).

Table 3.10 Association analysis of low-frequency variants in single genes with schizophrenia

	CMC	WSS	RANK	DBM*	REP*	SKAT	C-alpha	MAF
PRKAB2 [a]	1.000	0.874	0.713	0.507	0.729	0.887	0.768	0.05
FMO5	1.000	0.674	0.870	NA	0.679	0.620	1.000	0.01
	0.744	0.696	0.649	1.000	0.365	0.429	1.000	0.03
	1.000	0.922	0.711	0.883	0.584	0.647	0.672	0.05
CHD1L	0.282	0.233	0.181	0.136	0.096	0.541	1.000	0.01
	0.728	0.834	0.296	**0.029**	0.297	0.156	0.107	0.03
	0.506	0.724	0.298	**0.043**	0.373	0.188	0.303	0.05
BCL9	0.525	0.483	0.235	0.172	0.256	0.372	1.000	0.01
	1.000	0.942	0.256	0.143	0.353	0.260	0.744	0.03
	1.000	0.984	0.276	0.211	0.400	0.500	0.386	0.05
ACP6	0.545	0.494	0.503	0.932	0.329	0.696	0.757	0.01
	0.832	0.772	0.574	0.255	0.550	0.509	1.000	0.03
	0.388	0.491	0.568	0.302	0.315	0.422	0.894	0.05
GJA8 [a]	1.000	0.951	0.477	0.512	0.361	0.316	1.000	0.01
	0.805	0.670	0.565	0.159	0.660	0.429	0.751	0.05

[a]In the *PRKAB2* gene only a single variant at the MAF cutoff of 1% was identified which was a singleton observation. The remaining variants at *PRKAB2* belonged to the 5% MAF cutoff. In the *GJA8* gene no variants in the MAF cutoff of 3% were observed.
Analysis was not performed for *GJA5* gene since there was only one control specific variant observed in this gene.
The reported p-values for *CHD1L* are given for calculations based on 10000 permutations and for the other genes they are given for calculations based on 1000 permutations as the analyses were not re-ran with 10000 permutations in the absence of any significant p-values. Significant p-values below 5% significance level are given in bold.
*All the reported p-values from the DBM and REP are given for the two- and one-sided test statistics, respectively, which have performed better in the *CHD1L* gene.

A significant association signal was detected only in the *CHD1L* gene at both 3 and 5% MAF cut-offs by the DBM method with p-values (*P*=0.029 and 0.043, respectively) similar to what was obtained from the whole region analysis (Table 3.10). The analysis results of the remaining genes revealed no p-value below a significance threshold of 10% at none of the MAF cut-offs suggesting that *CHD1L* is the sole signal driving gene in the region. The significance in this dataset was reported by the two-sided test statistic of DBM which has a weighting scheme based on the allelic distribution of variants in the cases. In accordance, a denser spatial clustering can be seen for the variants detected in patients in comparison to the variants detected in controls in Figure 3.4. Although there should not necessarily be a single signal driver

cluster structure and none can be pinpointed by this method; there seems to be particular variant clusters around exons 10-14 and exons 16-19 in patients. The first cluster (exons 10-14) contains all the exons encoding for the Helicase C-terminal domain and second cluster (exons 16-19) contains two exons (exons 18 and 19) encoding for parts of the macrodomain of CHD1L protein.

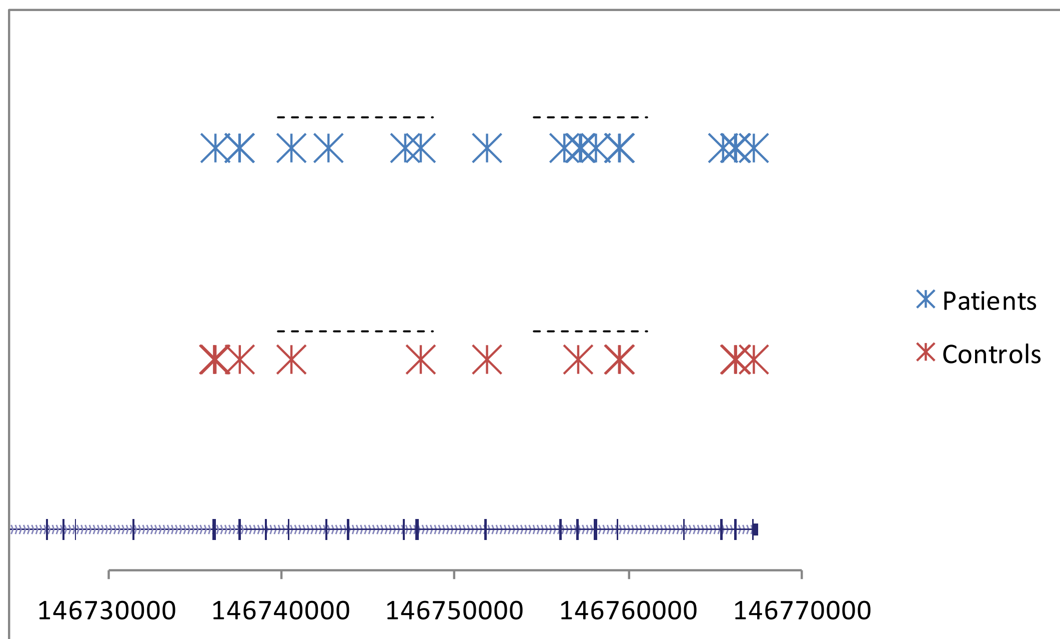


Figure 3.4 Spatial distribution of low-frequeny variants observed in the *CHD1L* gene. The spatial distribution of observed variants in patients (blue) and controls (red) are given regardless of the number or alleles detected. The x-axis denotes the genomic locations of the observed variants. The exon/intron structure of the *CHD1L* transcript 1 (NM_004284.3) given above the x-axis starts from intron 2 and ends with the last exon of *CHD1L* (exon 23) (left to right). Variant clusters around exons 10-14 and exons 16-19 are denoted with dashed lines.

3.2.3 Follow-up analyses of *CHD1L* gene

3.2.3.1 Genotyping and association analysis of the low-frequency *CHD1L* variants

In order to confirm the association signal observed in the *CHD1L* gene the low-frequency *CHD1L* variants were genotyped in an independent, large follow-up patient-control cohort of German origin. Except for one intronic deletion (chr1: 146756234) which could not be assayed, all of the low-frequency variants in the *CHD1L* gene were genotyped in an additional 1900 patients and 2186 controls of German origin (Genotyping-1 sample). In order to exclude that the ungenotyped variant was a significant driver of the initial association signal, the analysis of the resequencing data in *CHD1L* gene was repeated by removing this variant. The association signal remained robust (DBM, P=0.027 and 0.042 at MAF 3 and 5%, respectively, Attachment V). All of the genotyped variants yielded good quality cluster plots in the Sequenom genotyping assay. Sanger sequencing was performed for missing genotype calls to ensure that all the individuals have complete calls for the entire variant set. The analysis was performed by the application of

the outlined methods i) on the total sample (sequencing and genotyping samples together) and ii) only the genotyping sample, at the same MAF-cutoffs of 1, 3 and 5%. The results are depicted in Table 3.11. Although no significant association was observed by the DBM method, significant associations below 5% significance level were detected by the SKAT method at the 3% MAF cut-off both in the total sample (*P*=0.021) and the genotyping sample (*P*=0.028). Association analysis at the higher MAF cut-off of 5% yielded p-values below 10% significance level by SKAT in the total sample (*P*=0.062*)* and the genotyping sample (*P*=0.095*)*.

Table 3.11 Association analysis of low-frequency variants in *CHD1L* gene from genotyping

	MAF	CMC	WSS	RANK	DBM	REP	SKAT	C-alpha
Total sample	0.01	0.825	0.825	1.000	0.327	0.503	0.286	0.869
	0.03	0.310	0.269	0.117	0.462	0.249	**0.021**	0.212
	0.05	0.694	0.773	0.162	0.336	0.484	0.062	0.117
Genotyping sample	0.01	1.000	0.912	0.795	0.648	0.572	0.331	0.811
	0.03	0.520	0.471	0.219	0.753	0.237	**0.028**	0.219
	0.05	0.798	0.858	0.293	0.671	0.488	0.095	0.196

The reported p-values for all test statistics are given for calculations based on 10 000 permutations. Significant p-values below 5% significance level are given in bold. The better performing two- and one-sided test statistics for DBM and REP methods are given, respectively.

Next, a single marker association analysis was performed in the *CHD1L* gene in order to assess whether there were variants individually associated to schizophrenia with a risk or protective effect direction. For this purpose, Pearson's chi-square test was applied on individual variant data followed by random permutations on the case-control status to re-generate the test statistic for every permutated sample and to derive empirical p-values based on the generated distribution of the test statistics. The single marker analysis was applied both on the total sample and on only the genotyping sample (Attachment VI). The only variant which yielded a significant individual association to schizophrenia below 5% significance threshold was rs36008075 with an overrepresentation in patients (P_{total}=5x10^{-04}). The significance remained when the analysis was performed only on the genotyping sample ($P_{genotyping}$=9.5x10^{-04}) and also even after correction for multiple testing both in the total sample and the genotyping sample (P_{total}=0.01 and $P_{genotyping}$=0.02). A missense variant (I765M) located in the macro domain of CHD1L also yielded a nominal association below 10% significance threshold in the total and only genotyping samples (*P*= 0.080 and *P*=0.067, respectively). However, unlike rs36008075, this variant was overrepresented in control individuals implicating a protective effect direction.

3.2.3.1.1 Functional assessment of rs36008075

Splice site analysis

The marker which showed *per se* significant association to schizophrenia was an exonic variant which was located 57 bp downstream of the wild acceptor site of exon 7. This variant was predicted to create an alternative acceptor site with a consensus value (85.5) higher than that of the original acceptor site (73.9) by the HSF (Figure 3.5). If the prediction was correct, it was

expected that 57 bp of sequence information would not be transcribed into mRNA leading to skipping of 19 amino acid residues. In order to assess whether the predicted effect was true, first a Sanger sequencing was performed on genomic DNA of 280 individuals from whom lymphoblastoid cell cultures were also available.

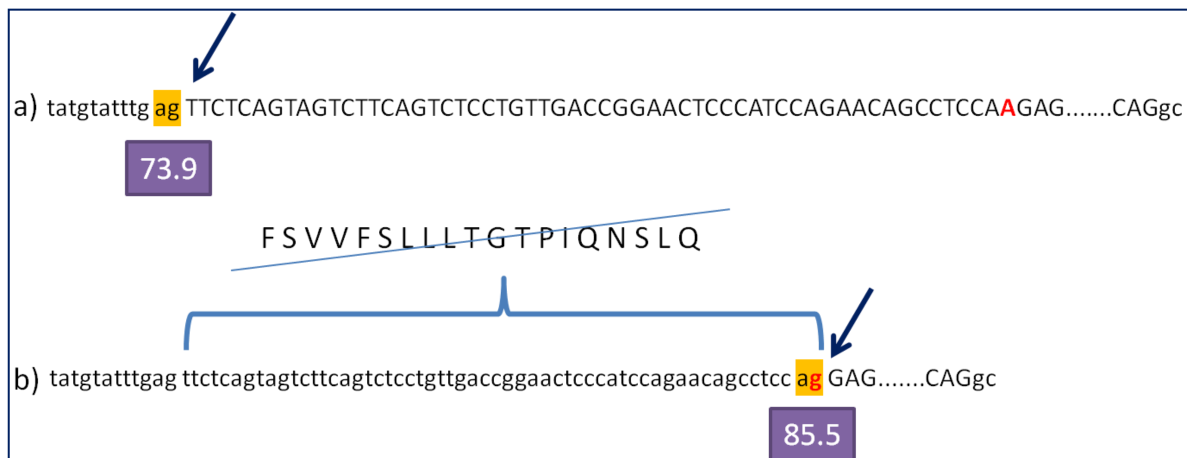


Figure 3.5 Exonic variant predicted to create a cryptic splice site by *in silico* assessment with Human Splicing Finder The intronic and exonic sequences are given lower and upper cases, respectively. The altered base is given in red. The (a) wild and (b) alternative acceptor sites are marked with yellow boxes and the first exonic bases after the wild and alternative splice sites are denoted with arrows. The consensus value of the (a) wild acceptor site motif is 73.9 and the consensus value of the (b) alternative acceptor site motif is 85.5. The amino acids (n=19) skipped due to the new acceptor site are given in the figure.

Data analysis identified five risk allele carrier individuals, all of which were heterozygous for the variant. As a next step, total RNA was isolated from the lymphoblastoid cell cultures of risk allele carriers (n=5) and individuals homozygous for the non-risk allele (n=15) to serve as controls. The mRNAs were reverse transcribed into cDNAs which then were Sanger sequenced for the region of interest (Figure 3.6). The analysis of the resequencing data revealed that the sequences which were predicted to have been skipped (57 bp) were indeed transcribed in the variant carrier individuals and that this variant did not lead to the predicted effect and the splicing occurred only at the true acceptor site (Figure 3.6).

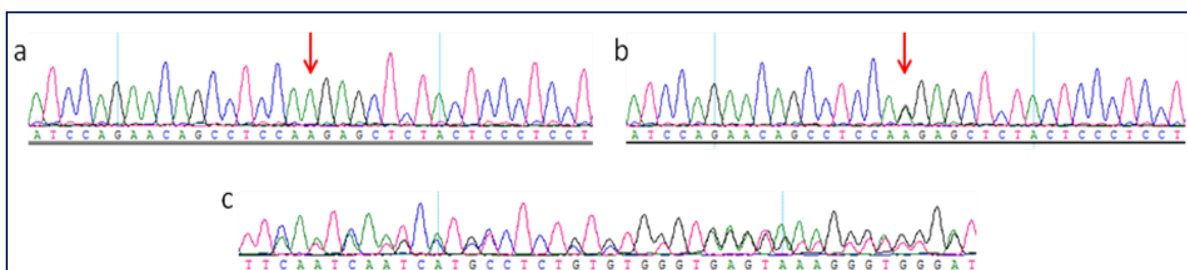


Figure 3.6 Electropherograms from cDNA sequencing in *CHD1L* gene. The electropherogram of an individual (a) homozygous for the non-risk allele and (b) heterozygous for rs36008075 are given where the point of variation is denoted with a red arrow. If the predicted splice site effect was true; two types of cDNA strands would have been available from a heterozygous rs36008075 carrier. The cDNA strands carrying the variant would have been shorter than the strands not carrying the variant by 57 bp and this would have mimicked the appreance of a heterozygous deletion in the electropherograms, such as the example electropherogram from a heterozygous deletion in genomic DNA from another region of *CHD1L* gene (c).

In order to take tissue specific effects into account the same analysis was performed on the cDNA reverse transcribed from total mRNA isolated from pre-mortem hippocampus tissue. For this purpose cDNAs of two risk allele carrier individuals and two individuals homozygous for the non-risk allele were assessed. One of the risk allele carriers was homozygous and the other was heterozygous for the variant. The results remained the same indicating that the predicted splice site effect does not occur in the hippocampus tissue either.

Expression analysis

Considering a possible regulatory role for rs36008075, the relative abundance of *CHD1L* transcripts were investigated in five individuals heterozygous for the risk allele (mean $age_{recruitment}$=49.0) and 15 individuals homozygous (mean $age_{recruitment}$=44.67) for the non-risk allele. For this purpose quantitative real-time PCR was performed on cDNA from the lymphoblastoid cell lines (Figure 3.7).

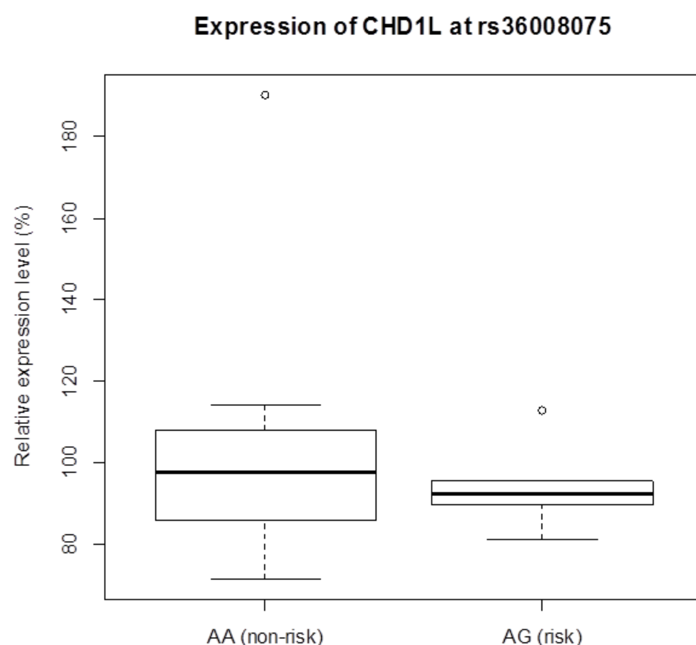


Figure 3.7 Boxplot for allele specific expression with respect to rs36008075. Boxplot for each genotype group (AA, n=15; AG, n=5) show the distribution of the expression values. The median values (middle black lines) and the quartiles (75%, top line; 25%, bottom line) are represented by the rectangles. The dashed vertical lines illustrate the distribution of the *CHD1L* expression levels. Outliers, showing more than 1.5 fold interquartile distance, are represented by a circle. The relative expression levels are calculated with reference to the mean expression value of non-risk allele carriers (AA) which is set as 100%. The relative mean expression value of risk allele carriers (AG) is 94.26%.

No significant difference between expression levels of risk allele carrier and non-carrier individuals was observed. However there was a non-significant reduction of *CHD1L* expression in risk allele carriers with about 5.7% difference in the mean expression values between the two groups (*P*=0.53, Welch two sample t-test) (Figure 3.7).

3.2.3.2 Extended resequencing of *CHD1L* gene and follow-up of individual variants

Extended resequencing of the *CHD1L* gene

Based on the evidence for involvement of low-frequency *CHD1L* variants in schizophrenia, the *CHD1L* gene was resequenced in an additional 96 patients to increase the power to identify

additional variants of potentially high penetrance risk effects (non-sense, splice site, frameshift etc...) which could mimic the effects of the 1q21.1 microdeletion. A total of six variants which were not detected in the initial resequencing step were detected in the extended resequencing of the *CHD1L* gene (Table 3.12).

Table 3.12 Low frequency variants identified in extended resequencing of *CHD1L* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Conservation[a]	Minor allele counts (n=patients)
146724404	rs143313938	C>T	intron 2	-	2	1
146728227	-	T>C	intron 5	-	5	1
146731457	-	A>C	intron 5	-	4	1
146737632	rs144288940	C>T	exon 8	R261*	5	1
146747766	rs113139670	A>G	acceptor site/ exon 14	-	5	1
146751866	-	T>G	donor site/ exon 15	-	5	1

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation). AA, amino acid; ss, splice site variant

Each of these variants was observed as a singleton event in different patients. Three of these were intronic variants (one known and two novel) with no direct effect on primary protein structure. The remaining three were defined as potentially functional variants. One of these was a non-sense variant and the other two were splice site variants located at essential acceptor and donor sites, respectively. The non-sense variant led to a premature stop codon in exon 8, at the amino acid position 261 (R261*) which is likely to lead to NMD although production of a truncated protein cannot be excluded (Figure 3.8)[182, 191]. One of the splice site variants (rs113139670) was located at the essential acceptor site of exon 14 and analysis by HSF suggested that the acceptor site is abolished by the consensus value changing from 88.3 to 59.35 (Attachment VII-A(i)). Analysis by HSF revealed two alternative acceptor sites. One of these is located 45 bp upstream of the abolished site with a consensus value of 85.23 and in case it is used as an alternative acceptor site it leads to insertion of 16 amino acids coupled to a frameshift event at position 462 (Attachment VII-A(ii)). The frameshift leads to a premature stop codon at amino acid position 485. The other alternative acceptor site is located 77 bp downstream of the abolished site with a consensus value of 84.27 and in case it is used as an alternative acceptor site it leads to deletion of 26 amino acids leading to a frameshift event at at position 462(Attachment VII-A(iii)). The frameshift leads to a premature stop codon at amino acid position 466. The second splice site variant was located at the essential donor site of exon 15 and was also suggested to lead to break down of the donor motif by the change of the consensus value from 87.66 to 60.83 (Attachment VII-B(i)). Analysis by the HSF revealed a cryptic donor site with a consensus value of 90.88 located 9 bp downstream of the broken site. If this cryptic donor site is used it leads to insertion of 4 amino acids coupled to a frameshift event at position 569 (Attachment VII-B(ii)). The frameshift eventually leads to a premature

stop codon at amino acid position 600. In both of the splice site variants exon skipping cannot be excluded either entirely or partially (affecting only some of the transcripts).

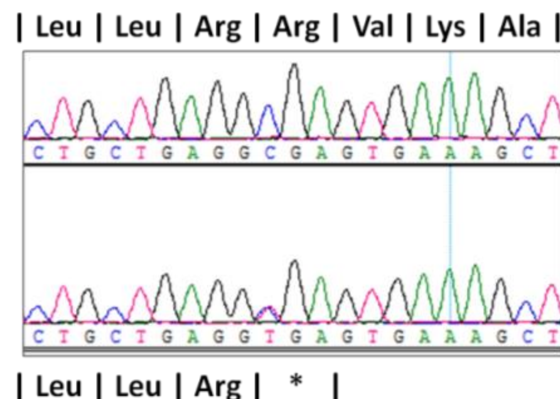


Figure 3.8 Non-sense variant observed in a schizophrenia patient in the *CHD1L* gene. The electropherogram of the wild type sequence (top) and the mutated sequence (bottom) are given. The triplet codons and the amino acids they encode are marked above and below of the wild and mutated sequences, respectively.

<u>Genotyping of potentially functional *CHD1L* variants</u>

As a next step the genetic association of individual *CHD1L* variants with schizophrenia was assessed by performing genotyping in a large patient-control cohort of German, Dutch and Danish origin (n=8768; Genotyping-2 sample). The aim was to target variants that are likely to have high penetrance effects contributing to increased disease risk. For this purpose potentially functional variants (missense, non-sense, splice site) identified in the initial or extended resequencing step which were observed in patients and in not more than one control individual were prioritized. From the total of seven low-frequency variants which fitted the selection criteria one failed (I827V) in the genotyping assay. Chi-square tests were applied followed by 1000 random permutations of case control status to derive empirical p-values. The empirical p-values were calculated both for the i) total sample (initial sequencing, extended sequencing and genotyping samples) and ii) the genotyping sample only (Table 3.13). The p-values were similar in both of the analyses and none of the individual markers yielded a significant association with schizophrenia in either of the analyses (Table 3.13). One of the splice site (rs113139670) variants and two missense variants (rs139791996, G700R; rs148289715, I765M) were equally distributed between patients and controls. The second splice site variant was not observed in additional patients but only in a single control individual. One patient specific missense variant from the initial resequencing sample (rs144757186, D381N) remained to be a singleton event after genotyping. The non-sense variant was observed in an additional four individuals. Three of these were patients of German (n=2) and Dutch (n=1) origin and one was a control individual of Danish origin leading to an overrepresentation in patients, yet not reaching statistical significance (P_{total}=0.138 and $P_{genotyping}$ = 0.245).

Table 3.13 Single marker association analysis of potentially functional variants in *CHD1L* gene

Genomic position (hg19)	dbSNP ID	Effect	Minor allele counts (n_{pat}/n_{cont})[a]	Total allele counts (n_{pat}/n_{cont})[a]	Empirical P-value[a]	Minor allele counts (n_{pat}/n_{cont})[b]	Total allele counts (n_{pat}/n_{cont})[b]	Empirical P-value[b]
146737632	rs144288940	R261*	4/1	9018/9072	0.138	3/1	8642/8884	0.245
146742648	rs144757186	D381N	1/0	8996/9052	0.245	0/0	8620/8864	-
146747766	rs113139670	splice site	20/19	9018/9068	0.798	19/19	8642/8880	0.935
146751866	-	splice site	1/1	9026/9074	0.752	0/1	8650/8886	0.748
146758054	rs139791996	G700R	45/43	9014/9060	0.786	42/43	8638/8872	0.953
146759387	rs148289715	I765M	19/27	9024/9070	0.271	18/26	8648/8882	0.282

Empirical p-values are calculated for the total ([a]) and follow-up ([b]) samples by 1000 times permutated case-control labels for Pearson's chi-square test. Minor and total allele counts of each individual variant are given separately based on the respective genotype calls.

The single marker analyses were also performed by application of logistic regression controlling for ethnicity (Attachment VIII). Empirical p-values were derived from 1000 permutations where the specific case/control ratios of every population were kept constant. No substantial differences between the p-values from Chi-square and logistic regression analyses were detected for any of the markers.

3.2.4 Follow-up analyses of *FMO5* gene

Extended resequencing of the *FMO5* gene

Although no statistical support from the initial resequencing analysis was received for *FMO5*, the identification of the only non-sense variant at this gene made it a gene of interest as non-sense variants are generally expected to have more deleterious effects and because this variant might have a similar functional effect as the 1q21.1 deletion. Therefore, first the *FMO5* gene was resequenced gene in an additional 96 patients to identify additional variants of potential interest. A total of 2 variants which were not detected in the initial resequencing step were detected in the extended resequencing of the *FMO5* gene (Table 3.14). Each of these variants was observed as a singleton event in different patients. One of these was a synonymous variant located at exon 2 with no direct effect on primary protein structure. The other variant was a missense variant leading to an arginine to glycine substititution (R319G) at a highly conserved region in exon 7, the substitution was predicted to be benign both by PolyPhen and SIFT.

Table 3.14 Low frequency variants identified in extended resequencing of *FMO5* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	AA changes	Conservation[a]	Effect (SIFT/PolyPhen)	Minor allele counts (n=patients)
146696583	-	C>G	exon 2	syn	-	2	1
146672962	rs142335408	T>C	exon 7	R319G	tolerated/benign	5	1

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation). AA, amino acid; syn, synonymous

Genotyping of potentially functional *FMO5* variants

As a next step, the functional variants identified in *FMO5* (in the initial or extended resequencing step) which were observed at least in one patient and in not more than one control individual were genotyped in order to identify risk variants of relatively high

penetrance. Genotyping was performed in the large patient-control cohort of German, Dutch and Danish origin (n=8768, Genotyping-2 sample). A total of 3 markers (missense and non-sense) fitted the selection criteria for genotyping. Chi-square tests were applied followed by 1000 random permutations of case control status to derive empirical p-values. The empirical p-values were calculated both for the i) total sample (initial sequencing, extended sequencing and genotyping samples) and ii) the genotyping sample only (Table 3.15). The analysis showed that the missense variant (rs142335408, R319G) identified in the extended resequencing step was significantly overrepresented in patients both in the total (P_{total}=0.032) and the follow-up ($P_{genotyping}$=0.048) samples (Table 3.15). The single marker analysis by logistic regression (controlling for the ethnicity) yielded similar p-values (P_{total}=0.032 and $P_{genotyping}$=0.053) (Attachment IX). The other genotyped missense variant (rs58351438, K166E) was non-significantly overrepresented in controls (P_{total}=0.270 and $P_{genotyping}$=0.167) and no additional individuals were identified to carry the non-sense variant (R485*) observed in a single patient in the initial resequencing step.

Table 3.15 Single marker association analysis of potentially functional variants in *FMO5* gene

Genomic position (hg19)	dbSNP ID	Effect	Minor allele counts (n_{pat}/n_{cont})[a]	Total allele counts (n_{pat}/n_{cont})[a]	Empirical P-value[a]	Minor allele counts (n_{pat}/n_{cont})[b]	Total allele counts (n_{pat}/n_{cont})[b]	Empirical P-value[b]
146684095	rs58351438	K166E	63/76	9020/9072	0.270	58/75	8644/8884	0.167
146672962	rs142335408	R319G	17/7	9022/9068	0.032	16/7	8646/8880	0.048
146658628	-	R485*	1/0	9024/9070	0.251	0/0	8648/8882	-

Empirical p-values are calculated for the total ([a]) and follow-up ([b]) samples by permutated case-control labels for Pearson's chi-square test. Significant p-values below 5% significance level are given in bold.The reported p-values for R319G are given for calculations based on 10000 permutations and for the other variants for calculations based on 1000 permutations as the analys es were re-ran with 10000 permutations only if significant p-values were derived from the initial analysis. Minor and total allele counts of each individual variant are given separately based on the respective genotype calls.

3.3 Resequencing of the *TCF4* gene

Variants detected in German schizophrenia patients

A total of about 7.2 kb genomic sequence per individual, totalling 1.37 Mb of sequence information in the whole sample was analyzed. The sequence analysis of *TCF4* in 190 schizophrenia patients revealed a total of 11 variants. Three of these were denoted as common (MAF≥5%) and eight were denoted as low-frequency (MAF<3%) variants. The frequencies of the common variants (rs1788027, rs6567210, rs8766) in the patients were not significantly different from their frequencies in the 379 European individuals of 1000 Genomes (P>0.05, data not shown). Among the eight low-frequency variants, three were defined as potentially functional (Table 3.16). Two of the potentially functional variants were novel missense variants (P156T, F211L) and one was a splice site variant (rs148658897). All of these variants were singleton observations in different patients. The missense variant leading to a Proline to Threonine substitution (P156T) was predicted to be damaging by SIFT with a low prediction confidence and possibly damaging by Polyphen. This variant was located at the N-terminal of

the bipartite nuclear localization signal (NLS) of TCF4 defined by Sepp *et al.*[174] (Figure 3.9). The other missense variant leading to a Phenylalanine to Leucine substitution (F211L) was predicted to be tolerated by SIFT and probably damaging by Polyphen. This variant was located between the AD1 and AD2 domains of TCF4 (Figure 3.9). The splice site variant (rs148658897) was located in the acceptor site of exon 3. The *in silico* analysis by HSF suggested an increased consensus value by ~9.2%, potentially leading to an increased splice site recognition. The remaining variants were synonymous or intronic variants with no direct effects on primary protein structure.

Variants detected in 379 European individuals from 1000 Genomes Project

The 1000 Genomes data from the 379 European individuals was analyzed for the same analysis frame used in the patients (18 coding exons and 50 bp up- and down-stream flanking intronic sequences). A total of 11 low-frequency variants (MAF<3%) were reported of which 3 were also detected in the present resequencing sample (rs35918540, rs143555588, rs76956936). Among the low-frequency variants from 1000 Genomes data, 2 were missense variants and the others were exonic or intronic synonymous variants with no direct effects on primary protein structure (Table 3.16). The missense variant leading to a Glycine to Arginine substitution (G452R) was predicted to be damaging by SIFT and probably damaging by Polyphen. The other missense variant leading to an Alanine to Valine substitution (A315V) was predicted to be damaging by SIFT and possibly damaging by Polyphen.

3.3.1 Association analysis of low frequency variants in *TCF4* gene

All of the low-frequency variants identified in 190 German schizophrenia cases and in 379 European individuals (n=16) were genotyped in an independent German cohort of ~1800 patients and ~2250 controls. All of the variants yielded clearly distinguishable cluster plots for the genotypes. After genotyping, four variants remained to be patient-specific which were all initially detected in the patient resequencing sample. Among them, were the two novel missense variants P156T and F211L which remained to be singleton observations (Table 3.16). The patient carrying P156T had an age of onset of 16 years. The patient had a brother who carried the same mutation and was also diagnosed with schizophrenia. However, as DNA samples from other family members were not available, a complete segregation analysis could not be conducted. The patient who carried F211L had an age of onset of 15 years. No parental DNA was available for this individual. The two other patient-specific variants were the splice site variant (rs148658897) which was observed in an additional patient in the genotyping step and an intronic variant (rs144346949) which remained to be a singleton observation (Table 3.16).

Table 3.16 Low-frequency variants observed in *TCF4* gene

Genomic position (hg19)	dbSNP ID	Alteration	Location	Source	AA changes	Effect (SIFT/PolyPhen)	Conservation[a]/ MAF[b]	Minor allele counts (n=patients/controls)	
								Sequencing	Genotyping
53252586	rs148658897	A>G	intron 2/ SS	Seq	-	-	5/0.003	1	1/0
53018138	-	G>T	exon 7	Seq	P156T	damaging*/possibly damaging	4/0.003	1	0/0
53017550	rs191953257	A>C	intron 8	1000 Gen	-	-	5/-	0	0/0
52946804	-	G>T	exon 9	Seq	F211L	tolerated/probably damaging	5/0.003	1	0/0
52942992	-	G>A	intron 9	1000 Gen	-	-	5/-	0	0/0
52942811	rs189454938	T>C	intron 10	1000 Gen	-	-	5/0.003	0	2/2
52928807	rs35918540	C>T	intron 11	Both	-	-	5/0.03	7	93/129
52928751	-	G>A	exon 12	Seq	syn	-	4/0.003	1	1/3
52928743	rs147445499	G>A	exon 12	1000 Gen	A315V	damaging/possibly damaging	5/0.003	0	9/7
52928685	rs144346949	A>G	intron 12	Seq	-	-	5/0.003	1	0/0
52928669	-	A>G	intron 12	1000 Gen	-	-	4/0.003	0	3/1
52927150	rs143555588	T>C	intron 13	Both	-	-	3/0.01	4	22/20
52901933	-	C>T	intron 15	1000 Gen	-	-	3/0.01	0	12/21
52901911	rs138570124	C>T	exon 16	1000 Gen	G452R	damaging/probably damaging	5/-	0	0/0
52901846	rs143944746	C>G	exon 16	1000 Gen	syn	-	4/0.003	0	10/8
52895549	rs76956936	C>T	exon 19	Both	syn	-	5/0.003	1	3/2

[a]Evolutionary conservation among human, chimp, rhesus, mouse and dog are presented as scores defining the number of species where the nucleotide is conserved (e.g. Score of 5 denotes full conservation).
[b]Variants are classified in the corresponding MAF categories based on their combined case/control frequencies in the total sample.
* Low prediction confidence was reported by SIFT for this variant. SS, splice site; Seq, resequencing sample; 1000 Gen, 1000 Genomes sample; AA, amino acid; syn, synonymous; MAF, minor allele frequency.

The patient who carried the splice site variant from the sequencing step had a disease age of onset at 18 years and had inherited the variant from the unaffected mother. No parental DNA was available for the second individual carrying the splice variant and the individual carrying the intronic variant. From the variants retrieved from the 1000 Genomes data, 3 were not observed in the present large German case-control sample including the missense variant G452R (rs138570124) (Table 3.16). The other missense variant reported in 1000 Genomes dataset (rs147445499, A315V) was identified both in patients and in controls of the German genotyping sample (Table 3.16). The missense variants observed in the German resequencing and/or genotyping samples and their localization on the protein domain structure of TCF4 are given in Figure 3.9.

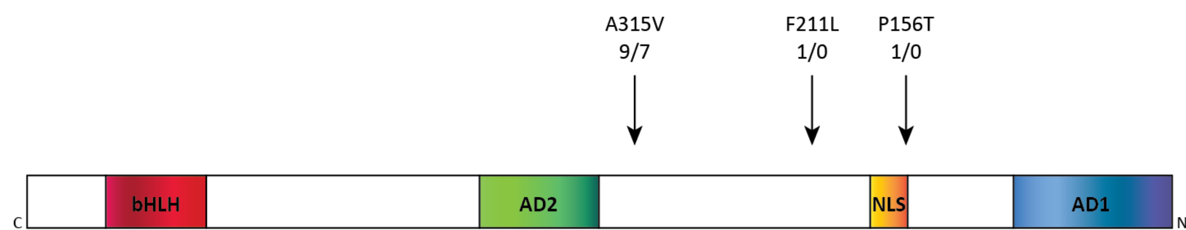


Figure 3.9 Protein domain structure of TCF4 and localization of observed rare missense variants. The protein domain structure is given for TCF4 isoform a (TCF4-B+) with 671 amino acid residues based on the study of Sepp *et al.*[174]. Only the variants which were observed in the German discovery and/or genotyping samples are depicted. Numbers of observations in the total sample are given for patients and controls ($n_{patients}/n_{controls}$). AD, transcription activation domain; NLS, nuclear localization signal; bHLH, basic helix-loop-helix domain.

The association analyses were performed by application of the CMC, WSS, RANK, DBM, REP, SKAT and C-alpha methods both on the total sample and only the genotyping sample at MAF cut-offs of 0.3, 1 and 3% (Table 3.17). The total sample consisted of the German patient discovery sample (n=190) and the independent German case-control cohort. It should be noted that the allele counts from 1000 Genomes individuals were not included in the total sample analysis to avoid introduction of possible sub-population specific rare variant patterns into the dataset (see section 2.6.8). The genotyping sample included only the independent German case-control cohort.

In the total sample analysis, when really rare variants with a MAF of <0.3% were considered, CMC, WSS, and REP revealed a significant association of *TCF4* with schizophrenia (*P*=0.045, 0.037 and 0.027, respectively) (Table 3.17).

Table 3.17 Association analysis of rare *TCF4* variants with schizophrenia

	MAF	CMC	WSS	RANK	DBM*	REP*	SKAT	C-alpha
Total sample	0.003	**0.045**	**0.037**	0.123	0.885	**0.027**	0.675	0.307
	0.010	0.136	0.118	0.128	0.775	**0.032**	0.302	0.981
	0.030	0.749	0.647	0.138	0.575	0.069	0.339	0.213
Genotyping sample	0.003	0.120	0.097	0.403	0.775	0.075	0.518	0.319
	0.010	0.272	0.220	0.336	0.750	0.094	0.465	0.808
	0.030	0.732	0.663	0.353	0.602	0.176	0.542	0.284

The reported p-values for all test statistics are calculated based on 10000 permutations. Significant p-values below 5% significance level are given in bold. *The better performing one-sided test statistic is given for the replication based strategy of Ionita-Laza *et al.*[178]

REP also reported a significant association at about the same level of significance for a MAF cut-off of 1% (*P*=0.032). When the analysis was restricted to the genotyping sample, the association signal of the region weakened and the p-values of all tested methods increased compared to the total sample. For even rarer variants with a MAF<0.3%, WSS and REP reported an association of *TCF4* region with schizophrenia below the 10% significance level (*P*=0.097 and 0.075, respectively), while the p-value from CMC was slightly above this level (*P*=0.120). For the higher MAF cut-off of 1%, REP revealed a p-value below the 10% significance level (*P*=0.094).

4. DISCUSSION

Recent genetic evidence suggests that disease loci can harbour different types of genetic variations[148, 149] ranging across the entire allele frequency/effect size spectrum. A plausible strategy for revealing the complete allelic disease spectrum is to resequence a locus identified previously through either GWASs or CNV studies since such an approach would capture rare and low-frequency sequence variants that are poorly represented on customized SNP chips. The generated sequencing data can then be utilized to assess the contribution of rare and low-frequency variants to disease susceptibility. Apart from this, in larger genomic regions associated with the disease, the presence of disease-associated variants may help to narrow down the region of interest. Potentially deleterious variants can be subjected to functional assays to obtain insight into pathophysiology of the particular disease. Based on this, sequencing analyses were performed in three schizophrenia regions identified through CNV studies or GWASs. These regions comprised of the *NRNX1* locus, 1q21.1 microdeletion region and the *TCF4* locus.

4.1 Rare genetic variation in the *NRXN1* gene and schizophrenia

Neurexins (NRXNs) are a family of presynaptic cell adhesion proteins encoded by three paralogous genes (*NRXN1, NRXN2, NRXN3*) each encoding a longer α-isoform and a shorter β-isoform which share identical intracellular sequences and differ in their extracellular sequences. The α-neurexins are transcribed from a promoter upstream of exon 1 whereas the β -neurexins are transcribed from an intragenic downstream promoter[192, 193]. Besides the alternative promoter usage, *NRXNs* are subjected to extensive alternative splicing which leads to generation of thousands of different NRXN isoforms[194]. Accumulating evidence[97, 98, 101, 103-107] has established large microdeletions disrupting the promoter region and the first exons of the *NRXN1* gene as a robust genetic risk factor for schizophrenia with the latest meta-analysis from Levinson *et al.* yielding an odds ratio estimate in the range of 7.5-8.2[45]. Common variants at the locus did not provide compelling evidence for association with schizophrenia in large GWASs published up to date[27, 29]. However, a putatively functional common variant was recently suggested to be associated with the response of European-American schizophrenia patients to clozapine; an atypical antipsychotic medication used in the treatment of schizophrenia[195].

Deletions at the *NRXN1* locus are one of the two known schizophrenia-associated CNVs which affect one single gene yielding direct and robust evidence for its involvement in disease aetiology. *NRXN1* is a biologically very plausible gene for schizophrenia as fundamental roles in synaptic differentiation, maturation and stabilization have been defined for the NRXN family both at inhibitory and excitatory synapses and mainly through their interaction with

postsynaptic neuroligins (NLGNs)[192, 193, 196] which were also previously associated with schizophrenia[197]. Besides these fundamental roles NRXN1- α is also shown to be essential for functioning of cortical voltage-gated Ca^{2+} channels and calcium (Ca^{2+})-triggered neurotransmitter release[198-200]. Homozygous neurexin-1α knock-out mice have been shown to exhibit behavioural deficits analogous to schizophrenia as well as selective impairment of excitatory neurotransmission at hippocampus[201].

Although the association of *NRXN1* microdeletions with schizophrenia is well-established, the possible contribution of rare *NRXN1* sequence variants to schizophrenia awaited further exploration. For this purpose a resequencing approach in the current study was applied at the locus in a sample consisting of 94 schizophrenia patients and 94 controls. Majority of the schizophrenia-associated *NRXN1* deletions affect the α-isoform and leave the β-isoform intact but some microdeletions affecting both isoforms have also been reported in schizophrenia patients[45, 97, 107]. Therefore the protein coding *NRXN1* exons were targeted for both of the two major NRXN1 isoforms (α2 and β) to investigate a potential contribution of small DNA alterations at the locus to schizophrenia. A total of 16 variants were identified at the locus which collectively were overrepresented in patients without reaching statistical significance. Possible explanations for this observation include insufficient statistical power of the investigated sample to detect very rare risk variants at statistical significance, a true negative finding (i.e. the observed overrepresentation has occurred by chance) or complex patterns of disease-associated and not associated variants at a locus that are difficult to dissect by a single statistical test. Statistical tests that show power advantages due to efficient combining of the signals of multiple rare variants across a genetic region for the association analysis of rare variants were either not available or were not conventionally in use by the time of the current study. Therefore the Fisher's exact test was applied to analyze the data which was one of the state-of-the-art methods by then[202]. However it is known that different statistical tests have variable sensitivies and powers to detect associations depending on the disease architectures at genetic loci[134, 203].

Among the 16 variants identified by resequencing, four were defined as potentially functional as they were leading to amino acid substitutions (E201G, P469S, L748I and I1068V). Two of these variants were exclusively observed in patients (E201G and I1068V) and they were followed up by genotyping to investigate potential high penetrance risk effects. No supportive evidence for such risk effects was found as one of these variants was identified in an additional control individual (E201G) and the other remained to be a singleton observation in the discovery patient (I1068V). The family data also did not yield any evidence for high penetrance effects. At the same time, it cannot be excluded that these are true risk variants for schizophrenia that are just very rare and thus difficult to detect reliably, even in relatively large sample sizes.

A number of independent publications have shown that *NRXN1* deletions appear to have pleiotropic effects and they do not only predispose to schizophrenia but also to a large spectrum of other neurodevelopmental phenotypes including autism spectrum disorders (ASD)[95, 196, 204-210]. The role of smaller *NRXN1* aberrations in autism had been addressed by prior sequencing studies in ASD cohorts. Evidence for overrepresentation of ultra rare missense and splice site variants in *NRXN1α*[211] as well as an overrepresentation of missense variants in *NRXN1β*[212] in autism patients has been presented. Another group also reported some *NRXN1* rare missense variants in their ASD cohort but did not find evidence for their overrepresentation in comparison to control cohorts[207]. Taking into account the possible pleiotropic effects of point *NRXN1* mutations in analogy to the *NRXN1* microdeletions, it was checked whether there was any overlap between the potentially functional variants identified in the current study and the ones reported by the aforementioned studies. Only the missense variant which was observed in 2 control individuals (L748I) in the present study was reported in ASD patients and controls by Kim *et al.*[207] and no supportive evidence was found for any of the two variants which were considered being most plausible for high penetrant risk effects (E201G and I1068V).

By the time these results were published[150] there was only one other study which had addressed the contribution of smaller rare variants in the *NRXN1* gene to schizophrenia[213]. However, based on the observation that *NRXN1* deletions mostly affect the promoter region and the first exons, the authors focused only on the promoter region which was not covered in our study. By resequencing the bidirectional *NRXN1α* promoter region in 170 patients and 160 controls they discovered two patient specific rare variants[213]. The *in silico* analysis suggested that these variants could exert allele-specific changes in binding of several transcription factors with some connections relevant to neuropsychiatric disorders. The results were however inconclusive since no genetic or experimental follow-up was reported to establishe a significant link between these variants and schizophrenia.

During the course of the present study, Gauthier *et al.*, reported their findings from exon targeted resequencing of *NRXN1*, *NRXN2* and *NRXN3* genes in patients suffering from schizophrenia (n=143) and other neurodevelopmental phenotypes (ASD=142; non-syndromic intellectual disability=94)[214]. The authors described a number of missense mutations they identified in their patients. No observations of E201G and I1068V were reported in this study either. However, P469S was identified in 2 ASD patients, 2 schizophrenia patients and 1 patient with non-syndromic intellectual disability and in none of the 190 controls. In the current study, this variant was detected in a patient as well as a control individual and therefore was not followed up. Based on the apparent overrepresentation of this variant in patients with neurodevelopmental phenotypes in the study of Gauthier at al.[214], it seems possible that this variant contributes to increased susceptibility to schizophrenia and/or other

neurodevelopmental diseases. Neverthless, genotyping in large samples would be necessary to establish these associations. Gauthier et al[214] identified L748I both in patients and in controls similar to Kim *et al.*[207] as described above. The combined evidence from us and the others therefore implicate no stronger risk or protective effect for this variant with regards to predisposition to neurodevelopmental phenotypes.

Although the current study failed to identify any, *NRXN1* point mutations of severe functional consequences and possible higher penetrance effects were reported in patients with schizophrenia by several studies which emerged after this study. The Gauthier *et al.*[214] study was indeed one of these. Within their schizophrenia cohort enriched for sporadic cases, the authors identified a patient with no family history of mental illness who carried a *de novo* heterozygous frameshift mutation in the *NRXN1* gene. The mutation led to a premature stop codon and a truncated protein which failed to localize at the cell surface, to promote synaptic differentiation and to bind two of its established postsynaptic binding partners. By exome sequencing Kong *et al.*,[215] also identified a *de novo* stop-gain mutation in *NRXN1* in a non-familial schizophrenic proband. Duong *et al.*, described a family with mutations in the *NRXN1* gene showing an intricate co-segregation with neurological, somatic and psychiatric disorders including schizophrenia[216]. The proband suffering from autism, non-syndromic intellectual disability and epilepsy was compound heterozygous for a 451 kb deletion affecting the promoter and the first four exons and a point mutation located at an essential splice site leading to a frameshift mutation and a premature stop codon. The deletion was inherited from a mother with sub-diagnostic autistic traits. The point mutation was inherited from the deceased father who was diagnosed with paranoid schizophrenia. The brother of the proband who suffered from psychotic disorder also carried only the point mutation. Interestingly all the above described point mutations observed in patients diagnosed with schizophrenia or psychotic disorder are leading to premature stop codons and therefore it is likely that they mimic the effects of *NRXN1* microdeletions. In the light of these findings it seems that not only CNVs but also gene disrupting point mutations in the *NRXN1* gene are predisposing to schizophrenia. Nevertheless, gene-based mining of large exome sequencing datasets in the near future are awaited to provide the complementary statistical support for the link between such point mutations and schizophrenia in analogy to the *NRXN1* microdeletions.

The current study failed to provide evidence for association of rare *NRXN1* variants with schizophrenia however it should not be overlooked that the employed approach suffered from certain limitations. One of these was related to the discovery sample size which had limited power to detect rare variants associations. The reason for the limited sample size was that *NRXN1* is one of the largest human genes[217] and Sanger sequencing of the gene in 188 individuals is a labour intensive and costly process even if only the coding sequences are

targeted. Secondly, for the discovery sample the aim was to select patients for whom parental DNA was available to check for *de novo* occurrence of an identified mutation and for co-segregation analysis. This prerequisite limited the ability to enrich the sample for patients with a higher likelihood of having (stronger) genetic causes for disease (e.g. patients with an early age of onset and/or positive family history of disease[218]) Eventually a very small fraction of the patient discovery sample (5 out of 94 patients) had an early age of onset and/or positive family history of schizophrenia (see section 2.5) limiting the chances for identification of highly penetrant mutations. A final limitation of this study can be the follow-up strategy which was restricted to genotyping of only patient specific potentially functional variants in an independent case-control cohort. First of all, by applying patient exclusiveness as a selection criterion, disease associations of variants with smaller effect sizes (that are also more likely to be detected in controls) might have been overlooked. For example this might be relevant for the missense variant P469S which is discussed above. Secondly, the majority of the identified sequence variants in our discovery cohort (n=12/16) were assigned as 'non-functional' meaning that they reside in non-coding flanking sequences or are synonymous substitutions with no direct effect on primary protein structure. However; it is known that synonymous and intronic variants may also be associated with human disease[219-223] and can exert functional consequences particularly on the regulatory level[223]. Interestingly, seven of the 'non-functional' variants in our study were exclusively observed in patients and one of them was even detected in three individuals. In comparison to this, there was only one control specific synonymous variant observed in a single individual. Therefore it is possible that some associations between these 'non-functional' variants and schizophrenia were overlooked by restricting the follow-up analysis only to the potentially functional variants. For example, one of the synonymous variants in our dataset which was not followed up was associated with autism in the Chinese Han population (rs2303298, p=3.45E-006; OR=2.152) following our study. This variant is a common variant in the Asian populations whereas it is rare in the European populations (1000 Genomes Project: 12% vs. 0.04% MAF, respectively). Based on the evidence for shared genetic risk factors between autism and schizophrenia it remains plausible that this variant is also associated with increased schizophrenia risk.

In conclusion, some protein truncating *de novo* or co-segregating point mutations in *NRXN1* have been identified in schizophrenia. Apart from them, no other low-frequency or rare variant with milder consequences has been defined to be associated with schizophrenia. However the lack of identification of such variants in *NRXN1* does not necessarily implicate their absence. First of all only a limited number of studies have addressed the role of small sized aberrations in *NRXN1* with relevance to schizophrenia one of which is our study[150, 213, 214]. Secondly none of

these studies –including ours- had sufficiently large sample sizes or more accurate methodologies (sample ascertainment, analysis method, follow-up, etc..) to generate statistical support for individual or collective disease association of such variants[224]. Therefore further research in this locus with relevance to low-frequency and rare variants is warranted. A next step might be the pooling of data from several resequencing studies on a collaborative basis to increase the sample size and thus the power to get a handle on the influence of rare variants at the *NRXN1* gene in the development of schizophrenia.

4.2 Rare genetic variation in the 1q21.1 microdeletion region and schizophrenia

In large genome-wide surveys we and others had identified rare microdeletions on chromosome 1q21.1 as strong genetic risk factors for schizophrenia[98, 108]. The deletions were reported to be carried by 0.23% of the schizophrenia patients in comparison to 0.02% of the healthy individuals[108]. Since then, these associations have been replicated by several other studies[45, 225, 226] and the latest meta-analysis from Levinson *et al.* yielded an odds ratio estimate in the range of 8.2-9.5[45]. At the molecular level the reported microdeletion was identified in two different extents: a shorter form spanning 1.35 Mb and a larger form spanning about 2.19 Mb that contains the shorter form within its boundaries[108]. The shorter form represents the minimally deleted region which suggests that the genes relevant for the disease aetiology may lie among the 10 Refseq genes spanned by the 1.35 Mb large deletions. The larger and the shorter microdeletion regions are both flanked by tandem segmental duplications which suggest non-allelic homologous recombination may be the mechanism responsible from the formation of these microdeletions[108]. The role of common variants in this region has been addressed by analysis of SNP array data of the large sample in the SGENE study (4,718 schizophrenia patients and 41,201 controls) to which we also contributed[108] and no evidence was found for association of any of the tagged SNPs with schizophrenia. Also no evidence has emerged from the large schizophrenia GWAS datasets for common variants in this region[27, 29].

It remained to be determined i) whether the 1q21.1 region harbours other types of disease associated low-frequency or rare variants in addition to the reported CNVs and ii) which of the affected genes are responsible from the pathogenic effect of these microdeletions. To investigate these issues dissecting a greater fraction of genetic variation at this particular locus was necessary. Accordingly, in the current study a resequencing approach was carried out in 94 schizophrenia patients and 94 healthy controls to discover small sized low-frequency variants below the resolution of array-based technologies. Only 7 of the 10 genes spanned by the shorter form of the microdeletion could be targeted and due to their location in segmental duplications it was not possible to target the remaining 3 genes (Attachment II). In order to investigate the

biological plausibility of the individual genes in the region, microarray based expression analysis was performed in pre-mortem human hippocampus tissue samples from epilepsy patients. Hippocampus is a relevant target tissue in schizophrenia as anatomical and functional alterations in patients with schizophrenia is well-established in this brain region[227].

In order to assess disease association of low-frequency and rare variants in the region, the sequencing data from schizophrenia patients and healthy controls was analyzed by application of statistical methods developed for locus based association analysis of rare variants. Several of the state-of-the-art association tests were applied simultaneously, as adviced[203] and practiced[228] elsewhere. This is mainly because these methods are suggested to have different power advantages to detect associations depending on the underlying genetic architecture of a disease locus[134, 203]. Simulation studies[203] have shown that different tests can outperform the others depending on interchangeable factors such as the proportion and number of causal and non-causal variants, directionality of effects, differential effect sizes and variant frequencies in the region[134]. Our analysis revealed significant association of low-frequency variants with schizophrenia in the region as detected by the DBM method at 3% and 5% MAF cut-offs. DBM method was developed based on the evidence supported hypothesis that variants of shared effect directions may tend to cluster physically in the same genetic regions. For example variants located in close proximity in the DNA sequence could also be located in the same functional protein domains and have similar impact on disease risk[181]. Accordingly, deleterious rare variants and protective rare variants would be expected to cluster together in different genomic regions, whereas neutral variants would be expected to be sparsely distributed throughout the region[181]. DBM method assesses such spatial distribution patterns between variants observed in cases and controls and tests whether there are significant differences between these two groups. In this study such a significant difference was identified in the 1q21.1 microdeletion region.

The second major aim of the present study was to pinpoint the genes underlying the pathogenic effect of the microdeletions which could consequently implicate certain biological pathways involved in the disease pathophysiology. Therefore, the association signal was tried to be refined to one or few of the targeted genes by applying the same statistical methods to the data on gene level. By gene-based analysis *CHD1L* was identified as the sole signal driver gene in the region with significant DBM p-values very similar to those obtained from the whole region analysis. *CHD1L* was one of the genes for which a moderate expression in the hippocampus was confirmed by the microarray based expression analysis. With the aim of replicating this association signal an independent genotyping of the low-frequency variants in the *CHD1L* gene was performed in a case-control cohort of ~4000 German individuals. The association was replicated at 3% MAF cut-off with significant p-values from the SKAT but not the DBM method.

Of note, at the 3% MAF cut-off, SKAT was one of the two best performing tests after DBM in the association analysis of the initial sequence data. The other test was the C-alpha which is based on similar grounds with SKAT[179]. The lack of replication by the DBM method in the independent follow-up sample was not unexpected for a targeted genotyping based follow-up strategy. The DBM method is based on spatial information and a genotyping based follow-up fails to add any new location information to the dataset. The genotyping in reasonable sample sizes rather leads only to an increase in allele counts of pre-set variants and increases the odds that all or the vast majority of the variants would be observed at least once in each phenotype group (case vs. control). This is well anticipated in complex phenotypes since variants with full penetrance effects are not expected to be observed[27, 229]. Thus such a follow-up approach dilutes away patient and control specific spatial variant clusters leading to loss of an association signal detected by the DBM method applied on sequence data. The SKAT method assesses similarities between genotype patterns of individuals across multiple markers in a region of interest and tests whether there is a relation between genotypic similarities and phenotypic similarities (e.g. case-control status of individuals)[179]. Unlike the collapsing based burden tests (CMC, WSS, RANK, etc.), genomic similarity based methods like SKAT are robust to variable effect magnitudes, to presence of neutral variants and variants of different effects (risk and protective) and they are indeed suggested to perform better than burden tests in the presence of such mixed effect directions and neutral variants[179, 180, 203]. Indeed by the single marker association testing in the genotyping sample, some evidence was observed for presence of both risk and protective variants in the *CHD1L* gene. The variant which was identified to have possible protective effect was a missense variant (I765M) and showed association below 10% significance level (P=0.067). This variant had a MAF below 0.3% suggesting that larger sample sizes may be required to establish more significant disease associations[224]. The presence of protective rare variants in disease loci has been shown in other complex phenotypes[230, 231]. However in the *CHD1L* gene the presence of opposite effect directions and proportion of rare and protective disease associated variants should be addressed further by additional genetic and functional studies to refine the disease architecture at this locus as the initial evidence from the current study for this type of architecture is not very strong. The variant which showed risk effect in the current study was a synonymous variant (rs36008075) and showed significant association with schizophrenia below 5% significance level (P=9.5x10^{-04}). The functionality of this variant was investigated by i) testing *in silico* predicted splice site effect in lymphocytes and hippocampus and ii) assessing a possible regulatory role with allele specific expression analysis. While any splice site effects were ruled out, a non-significant decrease in *CHD1L* expression of individuals carrying the risk allele was observed. It should be kept in mind that only a small number of carrier individuals (n=5) could be included in the expression assay due to the rarity

of this variant. Therefore, it remains plausible that the observed decrease in expression in risk allele carriers is a true effect which cannot significantly be established with the small sample size employed. Assessment with larger samples is necessary in future. It is also possible that this variant has other functional consequences effecting mRNA stability, protein conformation and folding as such effects have been defined in some synonymous variants associated with human disease[220].

Two recent independent studies have reported on the effect of 1q21.1 copy number variants on the expression profiles of CNV spanned genes in i) EBV-transformed lymphoblastoid cell lines of individuals carrying 1q21.1 microdeletions or microduplications and showing a range of phenotypes including schizophrenia[232], and ii) dorsolateral prefrontal cortex of a schizophrenia patient carrying the 1q21.1 microdeletion[233]. Both studies have shown positive correlations between the expression levels of the majority of the spanned genes and the number of genomic copies of the 1q21.1 region. Interestingly in these studies *CHD1L* was either the top affected gene[232] or was one of the two top affected genes (with *BCL9*)[233] among all the genes lying in the region.

In order to identify additional mutations in the *CHD1L* gene with strong functional consequences, (e.g. a non-sense variant which can mimic the reduced expression effect of 1q21.1 microdeletions on *CHD1L*) an extended sequencing of the gene in additional 96 patients was performed. Indeed a non-sense mutation was identified in a single patient located at exon 8 (R261*). Although there are many exceptions it is generally anticipated that mRNAs harbouring PTCs located 50-55 bp upstream of the last exon-exon junction are subjected to degradation by NMD (50-55 nucleotide rule)[182, 191, 234]. Due to its position at an early exon it is therefore probable that R261* also leads to NMD. In the extended sequencing sample two splice site mutations in two different patients were identified which also might have strong functional consequences (i.e. leading to frameshift and PTCs as suggested by *in silico* analysis). As a next step, genotyping of several *CHD1L* variants was performed in a large case-control cohort of ~9000 individuals of German, Dutch and Danish origin to assess their individual disease associations. For this purpose a set of variants were prioritized for cost effectiveness. Therefore, selection criteria were tailored which would enable assessment of variants that are likely to have high penetrance risk effects. In accordance missense, non-sense and splice site variants detected in the initial or extended sequencing steps in patients and in not more one control individual were targeted. The genotyping did not reveal any significant disease associations for any of the screened variants. However it should be noted that 3 of the 6 genotyped variants remained to be very rare including a missense variant, a splice site variant and the non-sense variant R261* (MAF=0.006%, 0.01%, 0.028%, respectively). Therefore it can be argued that the present sample could have had limited power to detect individual disease associations of these

ultra rare variants, despite the fact that it comprised of several thousand individuals[224]. Interestingly the non-sense variant R261* was identified in additional 3 patients and in only 1 control individual. Taking that even the 1q21.1 microdeletions themselves are identified in healthy individuals it is well-anticipated that this variant also has incomplete penetrance[76, 98]. However the rarity of this variant necessitates the employment of even larger samples to obtain further genetic evidence and this might pose a significant challenge. One plausible approach is therefore to prove a biological outcome of this mutation that mimics the effects of the 1q21.1 microdeletions on *CHD1L*[232, 233]. Further functional studies assessing the relevance of this mutation for schizophrenia can then also be performed by using animal models or human cells derived from mutation carrier patients[148].

CHD1L encodes for the chromodomain helicase DNA binding protein 1-like which belongs to the SNF2 (sucrose non-fermenting 2)-like subfamily of the SNF2 superfamily of ATPases[235, 236]. CHD1L was shown to play roles in DNA repair[237, 238], chromatin remodelling[237, 239, 240], gene expression [241-243] and cell cycle[235, 243, 244] regulation. It is also known as *ALC1* (amplified in liver cancer 1) as it was originally identified as a candidate oncogene located in a human 1q21 region which is recurrently amplified in hepatocellular carcinomas[244]. Following reports firmly established the association between *CHD1L* and hepatocellular carcinoma[241, 245, 246]. Recently, *CHD1L* has also been implicated in congenital anomalies of the kidney and urinary tract[247], ovarian carcinoma[248] and colorectal carcinoma[235]. *CHD1L* is not the first cancer associated gene implicated in schizophrenia as several other ‘cancer genes’ have been related to schizophrenia in independent studies[249, 250] (e.g. Adenomatous polyposis coli gene, *APC*,[197]; Neuregulin-1[251]; Protein kinase B[252], etc..). In addition, a relationship between cancer and schizophrenia has been long recognized through epidemiological studies despite the presence of some contradictory reports [249, 250]. One explanation is that some genes involved in cancer proliferation might confer protective advantage in biological processes such as proliferation, migration and apoptosis that could have detrimental effects on neurodevelopment if dysregulated[249]. In *CHD1L*-transgenic mice CHD1L has been shown to promote cell proliferation by enhancing the G1/S phase transition in the cell cycle via regulating the expression of some cell cycle proteins (i.e. Cyclin A, Cyclin D1, Cyclin E, CDK2, CDK4, p21, Rb, $p27^{Kip1}$, and p53)[243, 244]. Some of these proteins (i.e. Cyclin A, Cyclin D1, Cyclin E, p21, $p27^{Kip1}$) have also been shown to be dysregulated in neural progenitor cells[253, 254], oligodendrocyte precursors[255], post-mitotic oligodendrocytes[256] and post-mitotic GABA cells[257] derived from schizophrenia patients and in some cases were shown to be accompanied by *in vitro* cell cycle abnormalities in patient derived cell cultures[254]. Alterations in cell cycle dynamics have been suggested to i) have downstream effects on neurogenesis if they affect neural progenitor cells[254, 258], ii) lead to abnormal cell cycle re-entry, de-differentiation and/or apoptosis in post-mitotic neural cells[256, 257]. Therefore the normal regulation of cell cycle

seems to be a very important component in control and maintenance of temporal and spatial cascade of neurodevelopment. Compellingly some of the proteins regulated by CHD1L are known to have other biological functions in different aspects of neurodevelopment other than cell-cycle regulation (e.g. $p27^{Kip1}$ and Rb in neuronal migration)[259].

CHD1L has also been shown to regulate ARHGEF9-mediated Cdc42 activation by up-regulating the expression of *ARHGEF9*[242]. Interestingly *ARHGEF9* has been suggested as a candidate blood biomarker gene in schizophrenia based on increased expression levels during high hallucinations states[260]. *ARHGEF9* encodes for collybistin, a guanine nucleotide exchange factor (GEF) which is most abundantly expressed in brain with up-regulated expression particularly during major neuronal differentiation and synaptogenesis times and it is suggested to play a role in postmitotic neurons[261, 262]. Collybistin selectively activates the small GTPase Cdc42[263] which plays a fundamental role in dendritic spine formation[264-267]. *Cdc42* was associated with lower expression in the dorsolateral prefrontal cortex (DLPFC) of schizophrenia patients and this was suggested to contribute to the observation of reduced dendritic spine density in the DLPFC of subjects with schizophrenia[268, 269]. Collybistin binds also gephyrin which is a scaffolding protein at GABAergic synapses. By regulating gephyrin clustering at GABAergic synapses; collybistin is known to play a pivotal role in formation and maintenance of postsynaptic $GABA_A$ receptor clusters[270]. Altered cortical GABA neurotransmission associated with cognitive deficits in schizophrenia is well documented by several lines of evidence including altered densities and activities of $GABA_A$ receptors[21, 22]. Just recently the *GPHN* gene encoding for gephyrin was implicated in schizophrenia based on discovery of rare exonic *GPHN* microdeletions in unrelated individuals with different neurodevelopmental phenotypes including ASD and schizophrenia[271].

CHD1L has also been the first cellular protein identified to bind Nur77 and inhibit Nur77-mediated apoptosis[272]. Nur77 is an inducible transcription factor which belongs to the orphan NR4A subgroup of nuclear receptor superfamily[273]. The members of this subgroup are early response genes induced by various physiological stimuli (e.g. stress, cytokines, growth factors, neurotransmitters, physical stimuli, etc...) and involved in various biological processes including cell cycle regulation (and apoptosis), neurological disease, steroidogenesis, inflammation, carcinogenesis and atherogenesis[273]. In the CNS, NR4A genes including *Nur77* are suggested to have important roles in regulation and modulation of the dopaminergic system and are implicated in psychiatric disorders including schizophrenia[274] which has long been associated with imbalanced dopamine neurotransmission[19, 20]. *Nur77* expression was shown to be reduced in the prefrontal cortex of schizophrenia patients[275]. Administration of antipsychotic drugs which block dopamine D2 receptors are known to induce expression of *Nur77*[276, 277]. *Nur77* is suggested to regulate the expression of some dopamine neurotransmission genes[278]. Dopamine

neuron biochemical activity and dopamine turnover were shown to be altered in *Nur77* knockout mice[279].

Based on the presented findings from literatiure it is plausible that CHD1L contributes to the aetiology of schizophrenia through its contribution to several pathways implicated in neurodevelopment and brain functioning. However it should be noted that all the studies reporting the regulatory roles of CHD1L were conducted in the context of CHD1L's tumorigenesis promoting activities and with cells or tissues outside of the CNS[242-244, 272]. As *CHD1L* is expressed in human brain, further studies are awaited to confirm that the regulatory roles of CHD1L in CNS are overlapping with those outside of CNS.

The genetic overlap between schizophrenia and autism spectrum disorders is well established by the identification of pleiotropic CNVs[95]. Recent human genetic studies of autism have yielded some supportive evidence for the findings of the current work. First of all recurrent *de novo* mutations in autism have been identified in the chromodomain helicase binding protein 8 (*CHD8*) gene [280-282] which is related to the *CHD1L* gene as it encodes for a member of the protein family CHD1L belongs to. Second supportive evidence came from the study of Girirajan *et al.* [283] who performed a customized targeted microarray based CNV analysis in individuals with autism and healthy controls to refine the known ASD associated CNV regions to one or a few candidate genes. 1q21.1 region was also targeted in this study since 1q21.1 CNVs are among those which are predisposing both to schizophrenia and autism[225, 284, 285]. The authors identified a small atypical microdeletion spanning only the *CHD1L* gene in the 1q21.1 region in an individual with autism which was not observed in any of the controls[283]. Girirajan *et al.* therefore suggested *CHD1L* to be "a compelling candidate" gene for autism.

Due to the significance of non-sense mutations as described above some of the follow-up efforts in this study were also focused on to the *FMO5* gene which was the only gene where a non-sense mutation in a single patient from the discovery cohort was identified. The non-sense mutation was located at the terminal exon of *FMO5* and therefore it is probable that a truncated protein is formed by avoiding the NMD[182, 185, 191]. Yet it cannot be excluded that other NMD mechanisms which act irrespective of the "50-55 nucleotide rule"[234], such as non-sense mediated transcriptional gene silencing or non-sense mediated translational repression take place[191] to avoid the production of a truncated protein. *FMO5* was resequenced in an additional 96 patients in order to investigate whether the gene harboured more potentially functional risk variants, and data analysis revealed only another missense variant in this sample. Similar to the *CHD1L* gene, all the potentially functional *FMO5* variants observed in patients and in not more than one control were genotyped in the large genotyping sample of individuals with German, Dutch and Danish origin. The non-sense variant remained to be a singleton observation in the discovery sample after genotyping. Therefore it can neither be confirmed nor excluded that some

contribution of this variant to schizophrenia susceptibility exists. Interestingly, the genotyping identified one missense variant which was significantly associated with schizophrenia. FMO5 belongs to the family of flavin-containing monooxygenases (FMOs) which are microsomal enzymes responsible from oxygenative metabolism of several endogenous and exogenous chemicals[286]. Although the FMOs are primarily expressed in detoxication related tissues like liver, kidney and lung they were also shown to be expressed in mammalian brain[287, 288]. However the role of FMOs in CNS needs further exploration. In mice *FMO5* was one of the highest expression showing FMOs in brain[288, 289] and in humans *FMO5* expression was primarily localized to spinal cord with low expression levels in the remaining regions of the CNS[290].In the present study *FMO5* was one of the genes which yielded signal intensities below the background signal threshold in microarray analysis from hippocampus tissues. However, in the Allen Human Brain Atlas *FMO5* has been shown to be expressed in the human hippocampus. Therefore, it seems probable that low expression levels of *FMO5* fell behind the detection sensitivity of our microarray assay or that some of the tagging probes have failed in the microarray[291, 292]. Recent studies have implicated FMOs in Amyotrophic Lateral Sclerosis (ALS); an adult-onset, progressive, and fatal neurodegenerative disease[289, 290]. The link between FMOs and ALS was the suggested role of FMOs in detoxication under oxidative stress conditions based on their involvement in metabolism of nitrogen- and sulfur-groups present in proteins or chemicals[290]. While the role of FMOs and particularly FMO5 in CNS awaits further exploration, accumulating evidence suggests that oxidative stress may be involved in pathophysiology of schizophrenia[293]. Interestingly common variation in *FMO3* gene has recently been associated with the volume of lentiform nucleus; a bilateral structure in basal ganglia which is implicated in schizophrenia due to subtle volume alterations reported in schizophrenia and other psychiatric disorders[294]. However the extent of functional overlap between FMO3 and FMO5 is not yet clarified[188, 295]. Therefore, although our findings may implicate a possible contribution of rare *FMO5* variants to schizophrenia, their replication and further assessment of this gene are necessary since the overall functional evidence linkinging FMO5 to disease pathophysiology is currently weak.

It should be emphasized that our findings do not exclude the presence of other schizophrenia risk genes in the 1q21.1 region. Although no support for common variants in the region has emerged from large schizophrenia GWAS datasets, some independent studies reported association of common variants in *GJA8* and *BCL9* genes with schizophrenia. *GJA5* and *GJA8* encode for connexion 40 and connexion 50, respectively, which are subunits of gap junctions. Gap junctions are responsible from contacting plasma membranes by forming channel structures and are involved in direct metabolic and electrical communication among various cell types found in the mammalian brain[296]. They are also reported to be play crucial roles mammalian neural development[296, 297]. Prior to the discovery of the 1q21.1 microdeletions, the

chromosome 1q[298-300] and more specifically the 1q21-q22 region[301] had been implicated in schizophrenia by linkage analysis reports. Based on the their biological plausibility and their chromosomal location at the 1q21 region the role of *GJA5* and *GJA8* genes in schizophrenia had been addressed in a previous study before the discovery of the 1q21.1 microdeletions. The authors performed an association analysis of four polymorphisms for each of the two connexin genes in a sample of 190 schizophrenia patients and 190 controls of Caucasian origin. While the analysis revealed no association in the *GJA5* gene, a significant association between schizophrenia and a haplotype in the *GJA8* gene was identified. However no further reports linking the *GJA8* gene to schizophrenia have emerged since then, neither for common or rare variants. The current study also did not find any evidence for involvement of low-frequency or rare *GJA8* variants in schizophrenia. Moreover it is known that mutations in *GJA8* gene are causative of zonular pulverulent cataract (CZP1; OMIM: 116200) and cataract-microcornea syndrome (CAMIS; OMIM:116150]. Indeed one of the missense variants (I247V) identified in a patient in the present study sample was previously defined in a Russian mother and her son with zonular pulverulent cataract[189]. In the light of these the contribution of *GJA8* varians to schizophrenia remains to be a controversial issue. BCL9 has a critical role in the Wnt signal transduction cascade by promoting the transcriptional activity and nuclear retention of beta-catenin (β-catenin) [302-304]. The Wnt/ β-catenin signalling pathway has significant roles in various neurobiological domains both in developing and adult cental nervous system and converging evidence is supporting a role of this pathway in pathophysiology of several neuropsychiatric disorders including schizophrenia[305-308]. In addition a number of genes encoding components of Wnt signalling pathway have been associated with schizophrenia susceptibility[251, 309-312]. Due to its role in Wnt signalling pathway and its location in the 1q21.1 microdeletion region a previous study has investigated the role of common *BCL9* polymorphisms in schizophrenia in Chinese Han population and identified several SNPs significantly associated with the disease[313]. However due to the known genetic heterogeneity of schizophrenia susceptibility loci across different ethnic populations[314], the reported associations await replication in European populations by fine mapping of the locus. Yet *BCL9* gained some additional support from a recent report of a European-American GWAS meta-analysis on negative symptoms of schizophrenia[315]. Although no genome-wide significant hits were identified, a number of SNPs showed association with negative symptoms of schizophrenia at $p<5\times10^{-5}$ and the study-wide top association signal originated from an intronic SNP in the *BCL9* gene (rs583583, P=6.00×10^{-7}). It should be noted that in the study of Ye *et al.*, *BCL9* was the second top down-regulated gene with *CHD1L* in the dorsolateral prefrontal cortex of a schizophrenia patient carrying the 1q21.1 microdeletion[233]. Although *BCL9* was shown to be expressed in hippocampus by the microarray analysis, no evidence for the schizophrenia association of low-

frequency *BCL9* variants was found in the current study. Therefore it could be hypothesized that in the case the gene is a true schizophrenia risk gene, the allelic spectrum of disease in the locus may be restricted to only common variants. Finally it is noteworthy to re-mention that the *GPR89B, GPR89C* and *NBPF11* genes lying also in the minimally deleted 1q21.1 region could not be targeted by sequencing due to the technical limitations. Therefore it cannot formally be excluded that other disease related variants are present in these genes. However, based on empirical evidence from this study *CHD1L*, and from other studies *BCL9*, appear to be the most compelling candidates in the 1q21.1 microdeletion region so far.

In the current study, resequencing of the targeted genes was also performed in two 1q21.1 microdeletion carrier schizophrenia patients in order to explore the potential mechanism of 1q21.1 microdeletions in exerting their pathogenic effects. However resequencing data from these 2 patients did not provide evidence for an unmasking of recessive variants as no rare alteration of a potential functional relevance was found in them[190].

The majority of the schizophrenia associated CNVs span multiple genes[27]. The attempts to pin-point true disease relevant genes in these regions by targeted sequencing have been limited to prioritized gene screening as labour intensive and financially demanding efforts are required to screen complete gene sets. One example is the study by Carroll *et al.*[316] where they prioritized two genes (*PAK2*, and *DLG1*) in the 3q29 microdeletion region which is also associated with schizophrenia and performed a sequencing analysis in patients and controls[316]. The authors did not find any evidence for association of rare variants in these genes with schizophrenia. To the best of our knowledge; this is the first study that systematically and in unbiased approach screened all the targetable genes in a multiple gene spanning, schizophrenia associated CNV region. By the systematic analysis of low-frequency variants in the 1q21.1 microdeletion region and the follow-up replication steps it was discovered that low-frequency variants in the *CHD1L* gene contribute to the allelic spectrum of schizophrenia which presents *CHD1L* as a novel candidate gene in schizophrenia.

4.3 Rare genetic variation in the *TCF4* gene and schizophrenia

TCF4 at chromosome 18q21.2 was one of the first loci that surpassed the genome-wide significance threshold in the first wave of GWASs that emerged from collaborative efforts of big consortia[76]. The signal originating from a common SNP (rs9960767) located in an intron of the *TCF4* gene[76] was followed by an independent second association signal captured from two intergenic SNPs (rs4309482[82], rs12966547[87]) in perfect LD with each other and located between the coiled-coil domain containing 68 (*CCDC68*) gene and the *TCF4* gene. Finally another association signal from an intronic common SNP in moderate LD with rs9960767 was also reported in the recent PGC study[87]. The repeatedly reported association of common variants at

the locus with schizophrenia has thus made *TCF4* one of the most robust schizophrenia risk genes.

The *TCF4* gene encodes for transcription factor 4 which is a member of the class I basic helix-loop-helix (bHLH) protein family. This family is known to recognize and bind the Ephrussi-box ("E-box") motif on the DNA. Homodimerization as well as its heterodimerization with other molecules are necessary for DNA binding properties of TCF4. Interestingly some of the binding partners of TCF4 -such as HASH1, Math1 and neuroD2- are known to have fundamental roles in neurodevelopment[80, 317, 318]. *TCF4* itself is also a very plausible gene for schizophrenia based on several lines of evidence. It has been shown that TCF4 is important for neurodevelopment through its role in maturation of oligodendrocyte progenitor cells[319] . *TCF4* is widely expressed in adult as well as in developing human central nervous system with particularly higher expression levels reported in some tissues such as hippocampus and neocortex known to be relevant for development of schizophrenia[320]. Previously it was shown that mice over expressing *TCF4* in their forebrains exhibit cognitive impairments and pre-pulse inhibition which are analogous to deficits in schizophrenia patients[81]. Accordingly a recent study has shown significantly increased levels of *TCF4* expression in patients diagnosed with a spectrum of psychotic disorders including schizophrenia and bipolar disorder[321]. The same study reported the association of a set of common *TCF4* variants with negative symptoms, cognitive impairments and cerebellar volume in schizophrenia patients. *TCF4* is also a validated target gene of miR-137 which yielded the top association signal in the recent PGC schizophrenia GWAS meta-analysis[87, 322].

Despite the repetitive evidence linking common variation at *TCF4* to schizophrenia, the possible influence of rare *TCF4* sequence variants on schizophrenia susceptibility remained to be further explored. In order to address this issue exon-targeted resequencing of *TCF4* in 190 German schizophrenia patients was performed. To account for the genetic variation at the locus in control populations the variant data from 379 European individuals of 1000 Genomes Project was used. Using the data of 1000 Genomes Project instead of resequencing ethnically matched German controls was an approach undertaken due to its cost effectiveness. However one should be aware that the employment of publicly available datasets in rare variant studies should always be considered with caution since rare variants typically exhibit different and stronger stratification than common variants and many are expected to be population specific[323-325]. In other words, it remains unknown how accurately the rare genetic variance at a locus in a specific population represents the rare genetic variance at the locus in a different population. For example the 1000 Genomes dataset includes 5 different populations within the European sample (CEU, GBR, FIN, TSI, IBS) and substantially different rare variant patterns were shown even in these closely related populations (e.g. IBS and FIN populations carried excesses of rare

variants)[154]. Accordingly the rare variants retrieved from 1000 Genomes are not likely to precisely represent the variance at the *TCF4* locus in the German population. Therefore based on the argument that the use of rare variant information from geographically or ancestrally unmatched control individuals could introduce certain bias to the datasets[180, 324], an association analysis was not applied on the German patient discovery set and the 1000 Genomes discovery set[180, 324]. Rather an approach was undertaken where all the variants (n=16) defined in either of the two datasets were genotyped in an additional independent case-control sample of German origin. The genotyping revealed that although a substantial proportion of the 1000 Genomes variants (n=8/11) in the *TCF4* locus were observed in the German population some were not represented at all (n=3/11). This could mean that these variants are either population specific rare variants or are private mutations which are not mutually exclusive.

The association analyses were performed on i) the total sample composed of the genotyping sample and the discovery sample of 190 schizophrenia patients and ii) only the independent case-control genotyping sample. In the total sample analysis the allele counts from the 1000 Genomes individuals were not included based on the same rationale introduced for not performing an association analysis among the discovery samples.

In the total sample analysis three different rare variant association methods reported association of rare *TCF4* variants (MAF≤0.3%) with schizophrenia below the 5% significance level. However it is possible that these results were affected by type I error based on the possible bias introduced to the analysis by inclusion of genotypes of only the patient discovery sample. The same bias had previously been argued to be negligible in the case of a negative association report as described in the study of Dwyer *et al.*, where mutation screening followed by association analysis of rare sequence variants in the *ZNF804A* gene was performed[123]. However this argument does not apply to the current study where significant disease associations were observed and therefore one should be cautious with interpretation of the strong association signal observed in the total sample analysis. Yet, the analysis of only the genotyping sample which is free of this bias also revealed some evidence for association of rare *TCF4* sequence variants with schizophrenia. The association was observed with p-values below 10% significance level yielding a weaker signal from that observed in the total sample analysis. The weakening of the association was an outcome of the exclusion of allele counts from i) three patient-specific variants (P156T, F211L, rs144346949) which were observed only in the discovery sample; and ii) two other variants (rs148658897, rs76956936) for which patient allele counts were disproportionally higher in the smaller discovery sample. Observation of some association even after the loss of the contribution of these variants suggests that rare genetic variation at the *TCF4* locus might contribute to the allelic spectrum of schizophrenia risk. Just recently and after the completion of the current study Hu *et al.* reported deep

sequencing results from a number of schizophrenia genes identified through GWASs and showed an increased frequency of rare variants in cases compared with controls within a number of genes, one of which was the *TCF4* gene[326]. This finding is in line with the findings of the current study and when taken together it could be proposed that only common but also rare genetic variation at the *TCF4* locus may be involved in genetic susceptibility to schizophrenia.

In the present work, data analysis was performed by application of several methods from classes of statistical tests that can be roughly divided into collapsing and non-collapsing based association tests. All three methods which reported an association signal from the *TCF4* locus were collapsing based which is indicative of a shared effect direction by the majority of the causal variants in this study[179, 180, 203]. The REP method which always yielded the lowest p-values calculates two one-sided test statistics assessing risk and protective effects individually[178]. Since association signals were received from the one-sided test statistic of REP accounting for risk effects, it can be concluded that the major association signal in this region comes from risk variants. It should also be noted that the association signals were obtained at the lowest MAF cut-off of 0.3%. Recently Need *et al.* published their findings from exome sequencing of a schizophrenia case-control sample followed up by large scale genotyping of prioritized variants[145]. The major implication of this study is that there is a limited role for moderately rare risk variants (MAF=0.01-0.05) with moderate effect sizes and the contribution of rare variants to schizophrenia is more likely to arise from allelic heterogeneity with multiple variants of much lower frequencies[145]. The results of the current study are therefore in line with this proposition considering the *TCF4* locus where only truly rare variants (MAF<0.3%) seem to yield an association signal. However one cannot extend this interpretation to other disease loci as moderately rare variants with mild effect sizes can still be present and establishment of the disease associations of these variants would require much larger sample sizes and/or application of methodologies (e.g. collapsing based association tests) other than those employed in the Need *et al.*[145] study.

One of the schizophrenia GWAS signals implicating *TCF4* was located in an intergenic region between the *TCF4* and the *CCDC68* genes[82, 87]. *TCF4* was mostly assumed as the signal driving gene in the region based on the location of the other schizophrenia associated SNPs in the *TCF4* locus[76, 87, 320]. Yet this assumption awaited further genetic evidence for confirmation. Our findings might also serve this purpose as identification of locus specific disease associated rare variants is a plausible strategy for fine mapping of GWAS signals. This is mainly because the SNPs in commercial microarrays represent the genetic information from neighbouring markers due to LD structures that can span even several genes and therefore the disease associated common variants from GWASs are not necessarily the causal variants[71].

In the current study two novel patient-specific missense variants were identified which might play a substantial role in disease etiology. It is known that in addition to nonsense mutations and deletions leading to haploinsufficieny, there are a number of missense variants in the *TCF4* gene which are causative of Pitt Hopkins syndrome (PTHS; OMIM: 610954). PTHS is a severe neurodevelopmental disorder characterized by intellectual disability, facial dysmorphisms, breathing problems, postnatal microcephaly and epileptic seizures[327]. Apart from two exceptional variants located in the AD2 domain and between the bHLH and AD2 domains, all of the PTHS causative missense variants are accumulated at the bHLH domain responsible for DNA binding as well as homo- and heterodimerization properties of TCF4[328]. On the contrary, all of the missense variants observed in this study were located upstream of AD2 domain and are quite distant to the bHLH domain (Figure 3.9). Hu *et al.* reported the same observation, and found that in contrast to the PTHS mutations congregating in the C-terminal domains, the functional mutations identified in their cohorts were located principally in the N-terminal domains[326]. The distinctive spatial distribution of the missense variants in the present study might indicate different and possibly less severe functional consequences in comparison to the PTHS mutations. This would be in line with some previous arguments that while variants leading to severe TCF4 dysfunction could be causative of PTHS, variants with more subtle consequences could predispose to neuropsychiatric disorders by leading to alterations in several neuronal networks related to neurodevelopment (e.g. cognition, behavior, brain imaging endophenotypes)[81, 95, 321, 329]. Given previous reports of increased *TCF4* expression in schizophrenia patients[321] and schizophrenia-like deficits in transgenic mice with *TCF4* overexpression in the brain[81], a plausible hypothesis is that the missense mutations identified in the present and previous studies [326] of schizophrenia are gain-of-function mutations which lead to functionally similar effects with increased *TCF4* expression levels. This hypothesis is supported by several reports of gain-of-function mutations in transcription factors associated with other human diseases[330-332]. The different spatial distribution patterns of PTHS-causative and schizophrenia associated *TCF4* variants is an intriguing subject which warrants further investigation in functional studies. Such studies may elucidate the individual role of these missense variants and explore the relationship between the mutational *TCF4* spectrum and distinct neurobiological disorders.

As the effective application of exome or whole genome sequencing based study designs in complex phenotypes is still challenged by analytical and technical limitations[115, 134] one plausible approach is to follow up risk genes from GWASs by targeted sequencing. This approach has revealed the collective contribution of rare and common variants at individual loci to disease risk in other complex phenotypes[228, 230, 333, 334] and in some cases yielded important functional evidence. In the case of schizophrenia a handful of similar studies can be

mentioned[121, 123, 126, 335] and to the best of our knowledge the current study is one of the few to yield suggestive evidence with some statistical support that not only common but also multiple rare SNVs could be associated with schizophrenia at the same locus. Yet, one should be aware that these findings await confirmation mainly due to the limitation arising from the lack of rare variance information at the *TCF4* locus in an ethnically matched control sample. For this purpose, the best approach would be mining of large ethnically matched case-control datasets from next generation sequencing studies. This would also enable i) further exploration of the extent of contribution of rare variation at the locus to schizophrenia susceptibility and ii) assessment of possible interactive effects between rare and common *TCF4* risk variants.

To conclude, the better delineation of allelic mutation spectrum of schizophrenia in pre-defined GWAS loci would not only help filling in the missing heritability gaps, but would also lead to identification of rare disease associated variants of functional significance such as the missense variants discovered in this study. This would enable us to design functional studies which would provide insight into the specific alterations in molecular pathways leading to the development of schizophrenia.

5. SUMMARY

Schizophrenia is a highly heritable, multifactorial mental disorder. Up to date, a number of common SNPs with low penetrance and large rare structural variants (CNVs) with higher penetrance have been identified to contribute to schizophrenia susceptibility[27, 29]. The existence of highly penetrant rare CNVs demonstrates that rare variants account for part of the genetic background of schizophrenia[27]. This is also in accordance with a polygenic disease model that has been postulated for schizophrenia and which consists of risk variants ranging across the entire allele frequency/effect size spectrum[39]. The risk variants identified so far explain only part of the observed heritability, suggesting that many susceptibility variants, both common and rare, still await identification. Substantial efforts are currently made to find these risk variants in order to gain a more comprehensive understanding of the biological processes underlying schizophrenia. One promising strategy to identify so far unknown risk variants is to resequence genomic regions or genes that have already been highlighted through the identification of common risk SNPs[228, 230, 333, 334] or rare risk CNVs[148]. The underlying rationale is that different types of genetic variations can contribute to disease susceptibility at an individual locus [148, 149].

In the present doctoral thesis, sequencing analyses in three schizophrenia regions identified through CNV analyses or GWASs are described. One of these regions is located at 2p16.3 and contains the *NRXN1* gene. CNVs (in particular microdeletions) of different size affecting *NRXN1* had previously been identified at much higher frequency in schizophrenia patients than in controls[45]. Resequencing the coding exons and flanking splice sites of *NRXN1* in 94 schizophrenia patients and 94 controls revealed several low-frequency variants. However, data analysis did not identify a collective overrepresentation of these variants in schizophrenia patients. The functional variants that were exclusively observed in patients in this sample were then genetically followed up by genotyping in a larger case-control cohort. Analysis of the data did not identify any evidence for the individual or collective contribution of the genotyped variants to schizophrenia susceptibility either. In conclusion, the results of the described study did not generate any supportive evidence for the presence of schizophrenia associated rare and low-frequency variants in the *NRXN1* locus.

The second project described in this thesis was focused on the resequencing analysis of seven genes comprised in a microdeletion in 1q21.1 which had previously been described to confer a relatively high risk of developing schizophrenia[108]. By sequencing 94 schizophrenia patients and 94 controls, a total of 55 low-frequency variants in the 7 targeted genes were identified. Analysis of the data by several statistical tests revealed a significant overrepresentation of low-frequency variants (MAF<3%) in schizophrenia patients in comparison to healthy controls (P= 0.021). This suggests that not only the previously described risk CNV but also rare small sized

sequence variants in this region contribute to schizophrenia susceptibility. Another aim of this study was to further characterize the distribution and location of rare risk variants within the genomic region covered by the microdeletion to find out the gene/genes that are most likely to be the disease-relevant ones within the region. For this purpose the same statistical tests were applied on the sequencing data at the individual gene level. The results of this analysis identified *CHD1L* as the sole signal driver gene in the region (P=0.029, MAF<3%). In order to replicate the association finding in this gene all of the *CHD1L* variants were genotyped in a case-control cohort of 4086 German individuals. Analysis of the genotyping data confirmed the association between the low-frequency *CHD1L* variants and schizophrenia (P= 0.028, MAF<3%). Single marker analysis of the genotyped variants revealed one synonymous variant which was alone associated with disease in the genotyping sample (rs36008075, $P=9.5\times10^{-04}$). A regulatory effect for rs36008075 could neither be confirmed nor excluded by quantitative expression analysis, mainly due to the small sample size dictated by the low frequency of the minor allele. The mechanism through which rs36008075 contributes to disease susceptibility therefore awaits further identification. As a next step, extended sequencing of *CHD1L* was performed in additional 96 schizophrenia patients to search for variants of likely functional relevance that can mimic the 1q21.1 microdeletions. Data analysis revealed a non-sense mutation (rs144288940, R261*) which was further detected in 3 additional patients and in a single control individual by genotyping in a sample of 8768 individuals (P=0.245). The frequency of R261* needs to be monitored now in large samples of patients and controls to confirm its influence on disease susceptibility.

The biological plausibility of *CHD1L* as a schizophrenia susceptibility gene was supported by the gene expression analysis performed in pre-mortem human hippocampus tissue samples which demonstrated that *CHD1L* is expressed in human hippocampus, a brain region known to be affected in schizophrenia patients. In addition, literature mining revealed several studies describing the biological roles of CHD1L that could be relevant for schizophrenia and others suggesting *CHD1L* as a potential candidate gene for autism which is known to share a genetic overlap with schizophrenia. Taken together the results obtained in the course of this thesis strongly suggest that *CHD1L* is the most plausible risk gene for schizophrenia among the genes located in the microdeletion region.

The last project described in this thesis was focused on the *TCF4* gene which harbours schizophrenia-associated common variants. Resequencing of the *TCF4* gene in a total of 190 schizophrenia patients revealed 8 low-frequency variants (MAF<3%). In order to account for the genetic variation in control individuals, the low-frequency variant information at the *TCF4* locus was retrieved from 379 European individuals of the 1000 Genomes project. All of the variants identified in either of these datasets were then genotyped in a case-control cohort of

4069 German individuals. Application of several statistical analysis methods to analyze the data revealed a non-significant trend towards association (*P*=0.075, MAF<0.3%) suggesting that rare variants in this locus might contribute to disease susceptibility. Interestingly, in this study two ultra-rare, patient specific, novel missense variants were identified which could be the target of functional studies for better understanding of pathophysiological mechanisms underlying schizophrenia.

All together the studies described in this thesis show that different types of genetic variation can contribute to disease susceptibility at an individual locus and that targeted sequencing strategies in previously defined disease loci can lead to a better delineation of the allelic mutation spectrum of schizophrenia as well as to identification of the putative disease-associated genes in genomic candidate regions containing several genes. These investigations represent important steps on the long way from identifying disease-associated risk variants (SNPs, CNVs) towards understanding the underlying biological processes they are tagging.

6. OUTLOOK

The present thesis describes the resequencing analyses in three genomic regions that had previously been implicated in schizophrenia by multiple GWASs or CNV studies. The targeted sequencing of the selected regions was performed by Sanger sequencing since this was the standard technology by the start and during the course of the hereby described studies. With the continuously ongoing technological improvements, Sanger sequencing as a relatively expensive and slow technology is currently being replaced by the new generation sequencing technologies. These technologies enable much faster sequencing of many more patients only at a fraction of the costs for Sanger sequencing. The ability to sequence larger number of individuals in turn leads to increased power for detection of rare variants. Another asset of new generation sequencing technologies is that they would allow the complete coverage of larger candidate regions, such as the complete 1q21.1 microdeletion region (~1.35 Mb) or the complete genomic region spanned by the NRXN1 gene (~1.1 Mb). This will not only provide insights into rare variants that are located in exonic and splice consensus sites but also to those located at the intons or at regulatory gene regions. The necessity for assessment of such variants in terms of contribution to disease susceptibility is a very valid argument as results from GWASs in common diseases have shown that the disease-associated SNPs are more likely to be located in regulatory regions of genes[223].

Besides benefiting from the new generation sequencing technologies, it would be necessary in the near future to establish collaborations with other research groups for sharing of the sequencing data to further increase the samples sizes and thus the statistical power to detect rare variant associations. This has been successfully performed over the past few years by the Psychiatric Genomics Consortium with GWAS data sets for psychiatric disorders from dozens of research groups world-wide.

One of the major goals of sequencing studies is to deliver rare variants which are naturally expected to have a higher penetrance and thus more easily demonstrable functional consequences. Such variants can be assessed by functional assays for a better understing of the specific molecular alterations involved in the pathopyhsiology of schizophrenia, but functional investigation is complex and requires very specialized expertise depending on the nature and location of the associated variants. A closer collaboration between genetics research groups and groups who have an expertise in characterization of the functional outcomes of genetic variations by the use of cell culture and animal models is clearly warranted. Currently some collaborative attempts are ongoing regarding a selection of the rare potentially functional variants discovered in this thesis.

To conclude, a comprehensive understanding of the influence of rare variants to the overall heritability of schizophrenia and the exact nature of these variants will have profound effects on the understanding of the biology of this disease and may open avenues for new therapies. It is therefore of utmost importance that the strategy employed in the current thesis is further developed by all the aforementioned means

7. REFERENCES

1. McGrath J, Saha S, Chant D, Welham J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiol Rev 2008; 30: 67-76.

2. Tandon R, Nasrallah HA, Keshavan MS. Schizophrenia, "just the facts" 4. Clinical features and conceptualization. *Schizophr Res* 2009; 110(1-3): 1-23.

3. Keshavan MS, Nasrallah HA, Tandon R. Schizophrenia, "Just the Facts" 6. Moving ahead with the schizophrenia concept: from the elephant to the mouse. *Schizophr Res* 2011; 127(1-3): 3-13.

4. Robinson DG, Woerner MG, McMeniman M, Mendelowitz A, Bilder RM. Symptomatic and functional recovery from a first episode of schizophrenia or schizoaffective disorder. *Am J Psychiatry* 2004; 161(3): 473-479.

5. Harrison G, Hopper K, Craig T, Laska E, Siegel C, Wanderling J *et al.* Recovery from psychotic illness: a 15- and 25-year international follow-up study. *Br J Psychiatry* 2001; 178: 506-517.

6. Emsley R, Chiliza B, Asmal L, Harvey BH. The nature of relapse in schizophrenia. *BMC Psychiatry* 2013; 13: 50.

7. Tandon R, Nasrallah HA, Keshavan MS. Schizophrenia, "just the facts" 5. Treatment and prevention. Past, present, and future. *Schizophr Res* 2010; 122(1-3): 1-23.

8. Hyman SE. The diagnosis of mental disorders: the problem of reification. *Annu Rev Clin Psychol* 2010; 6: 155-179.

9. Shepherd AM, Laurens KR, Matheson SL, Carr VJ, Green MJ. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neurosci Biobehav Rev* 2012; 36(4): 1342-1356.

10. Fitzsimmons J, Kubicki M, Shenton ME. Review of functional and anatomical brain connectivity findings in schizophrenia. *Curr Opin Psychiatry* 2013; 26(2): 172-187.

11. Fornito A, Zalesky A, Pantelis C, Bullmore ET. Schizophrenia, neuroimaging and connectomics. *Neuroimage* 2012; 62(4): 2296-2314.

12. Whitfield-Gabrieli S, Ford JM. Default mode network activity and connectivity in psychopathology. *Annu Rev Clin Psychol* 2012; 8: 49-76.

13. Gur RE, Gur RC. Functional magnetic resonance imaging in schizophrenia. *Dialogues Clin Neurosci* 2010; 12(3): 333-343.

14. He Z, Deng W, Li M, Chen Z, Jiang L, Wang Q *et al.* Aberrant intrinsic brain activity and cognitive deficit in first-episode treatment-naive patients with schizophrenia. *Psychol Med* 2013; 43(4): 769-780.

15. Minzenberg MJ, Laird AR, Thelen S, Carter CS, Glahn DC. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch Gen Psychiatry* 2009; 66(8): 811-822.

16. Coyle JT, Balu D, Benneyworth M, Basu A, Roseman A. Beyond the dopamine receptor: novel therapeutic targets for treating schizophrenia. *Dialogues Clin Neurosci* 2010; 12(3): 359-382.

17. Kantrowitz J, Javitt DC. Glutamatergic transmission in schizophrenia: from basic research to clinical practice. *Curr Opin Psychiatry* 2012; 25(2): 96-102.

18. Coyle JT, Basu A, Benneyworth M, Balu D, Konopaske G. Glutamatergic synaptic dysregulation in schizophrenia: therapeutic implications. *Handb Exp Pharmacol* 2012; (213): 267-295.

19. Brunelin J, Fecteau S, Suaud-Chagny MF. Abnormal striatal dopamine transmission in schizophrenia. *Curr Med Chem* 2013; 20(3): 397-404.

20. Kuepper R, Skinbjerg M, Abi-Dargham A. The dopamine dysfunction in schizophrenia revisited: new insights into topography and course. *Handb Exp Pharmacol* 2012; (212): 1-26.

21. Gonzalez-Burgos G, Fish KN, Lewis DA. GABA neuron alterations, cortical circuit dysfunction and cognitive deficits in schizophrenia. *Neural Plast* 2011; 2011: 723184.

22. Stan AD, Lewis DA. Altered cortical GABA neurotransmission in schizophrenia: insights into novel therapeutic strategies. *Curr Pharm Biotechnol* 2012; 13(8): 1557-1562.

23. Silveira C, Marques-Teixeira J, de Bastos-Leite AJ. More than one century of schizophrenia: an evolving perspective. *J Nerv Ment Dis* 2012; 200(12): 1054-1057.

24. van Os J, Kenis G, Rutten BP. The environment and schizophrenia. *Nature* 2010; 468(7321): 203-212.

25. Lichtenstein P, Yip BH, Bjork C, Pawitan Y, Cannon TD, Sullivan PF *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* 2009; 373(9659): 234-239.

26. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* 2003; 60(12): 1187-1192.

27. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 2012; 13(8): 537-551.

28. Kim Y, Zerwas S, Trace SE, Sullivan PF. Schizophrenia genetics: where next? *Schizophr Bull* 2011; 37(3): 456-463.

29. Mowry BJ, Gratten J. The emerging spectrum of allelic variation in schizophrenia: current evidence and strategies for the identification and functional characterization of common and rare variants. *Mol Psychiatry* 2012.

30. Drew LJ, Crabtree GW, Markx S, Stark KL, Chaverneff F, Xu B *et al.* The 22q11.2 microdeletion: fifteen years of insights into the genetic and neural complexity of psychiatric disorders. *Int J Dev Neurosci* 2011; 29(3): 259-281.

31. Wray NR, Visscher PM. Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophr Bull* 2009; 36(1): 14-23.

32. Craddock N, O'Donovan MC, Owen MJ. Phenotypic and genetic complexity of psychosis. Invited commentary on Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 2007; 190: 200-203.

33. McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 2007; 190: 194-199.

34. Goldstein DB, Chikhi L. Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* 2002; 3: 129-152.

35. Lander ES. The new genomics: global views of biology. *Science* 1996; 274(5287): 536-539.

36. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001; 17(9): 502-510.

37. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet* 2012; 90(1): 7-24.

38. Rodriguez-Murillo L, Gogos JA, Karayiorgou M. The genetic architecture of schizophrenia: new mutations and emerging paradigms. *Annu Rev Med* 2012; 63: 63-80.

39. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; 460(7256): 748-752.

40. Gottesman, II, Shields J. A polygenic theory of schizophrenia. *Proc Natl Acad Sci U S A* 1967; 58(1): 199-205.

41. van Dongen J, Boomsma DI. The evolutionary paradox and the missing heritability of schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2013; 162B(2): 122-136.

42. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; 44(3): 247-250.

43. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001; 69(1): 124-137.

44. Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J *et al.* Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *Jama* 2006; 295(12): 1379-1388.

45. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 2011; 168(3): 302-316.

46. Doherty JL, O'Donovan MC, Owen MJ. Recent genomic advances in schizophrenia. *Clin Genet* 2012; 81(2): 103-109.

47. Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. *Hum Hered* 2011; 72(4): 228-236.

48. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003; 33 Suppl: 228-237.

49. Ng MY, Levinson DF, Faraone SV, Suarez BK, DeLisi LE, Arinami T *et al.* Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry* 2009; 14(8): 774-785.

50. MacIntyre DJ, Blackwood DH, Porteous DJ, Pickard BS, Muir WJ. Chromosomal abnormalities and mental illness. *Mol Psychiatry* 2003; 8(3): 275-287.

51. Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet* 2000; 9(9): 1415-1423.

52. Soares DC, Carlyle BC, Bradshaw NJ, Porteous DJ. DISC1: Structure, Function, and Therapeutic Potential for Major Mental Illness. *ACS Chem Neurosci* 2011; 2(11): 609-632.

53. Kamiya A, Sedlak TW, Pletnikov MV. DISC1 Pathway in Brain Development: Exploring Therapeutic Targets for Major Psychiatric Disorders. *Front Psychiatry* 2012; 3: 25.

54. Ishizuka K, Kamiya A, Oh EC, Kanki H, Seshadri S, Robinson JF *et al.* DISC1-dependent switch from progenitor proliferation to migration in the developing cortex. *Nature*; 473(7345): 92-96.

55. Bradshaw NJ, Porteous DJ. DISC1-binding proteins in neural development, signalling and schizophrenia. *Neuropharmacology* 2012; 62(3): 1230-1241.

56. Sanders AR, Duan J, Levinson DF, Shi J, He D, Hou C *et al.* No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. *Am J Psychiatry* 2008; 165(4): 497-506.

57. Kamnasaran D, Muir WJ, Ferguson-Smith MA, Cox DW. Disruption of the neuronal PAS3 gene in a family affected with schizophrenia. *J Med Genet* 2003; 40(5): 325-332.

58. Pickard BS, Malloy MP, Christoforou A, Thomson PA, Evans KL, Morris SW *et al.* Cytogenetic and genetic evidence supports a role for the kainate-type glutamate receptor gene, GRIK4, in schizophrenia and bipolar disorder. *Mol Psychiatry* 2006; 11(9): 847-857.

59. Knight HM, Pickard BS, Maclean A, Malloy MP, Soares DC, McRae AF *et al.* A cytogenetic abnormality and rare coding variants identify ABCA13 as a candidate gene in schizophrenia, bipolar disorder, and depression. *Am J Hum Genet* 2009; 85(6): 833-846.

60. Karayiorgou M, Simon TJ, Gogos JA. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. *Nat Rev Neurosci* 2010; 11(6): 402-416.

61. Pulver AE, Nestadt G, Goldberg R, Shprintzen RJ, Lamacz M, Wolyniec PS *et al.* Psychotic illness in patients diagnosed with velo-cardio-facial syndrome and their relatives. *J Nerv Ment Dis* 1994; 182(8): 476-478.

62. Murphy KC, Jones LA, Owen MJ. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch Gen Psychiatry* 1999; 56(10): 940-945.

63. Karayiorgou M, Morris MA, Morrow B, Shprintzen RJ, Goldberg R, Borrow J *et al.* Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc Natl Acad Sci U S A* 1995; 92(17): 7612-7616.

64. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 2008; 40(7): 880-885.

65. Rees E, Moskvina V, Owen MJ, O'Donovan MC, Kirov G. De novo rates and selection of schizophrenia-associated copy number variants. *Biol Psychiatry* 2011; 70(12): 1109-1114.

66. Bassett AS, Chow EW, AbdelMalik P, Gheorghiu M, Husted J, Weksberg R. The schizophrenia phenotype in 22q11 deletion syndrome. *Am J Psychiatry* 2003; 160(9): 1580-1586.

67. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G *et al.* Replicating genotype-phenotype associations. *Nature* 2007; 447(7145): 655-660.

68. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* 2012; 8(12): e1002822.

69. Consortium IH. The International HapMap Project. *Nature* 2003; 426(6968): 789-796.

70. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; 467(7311): 52-58.

71. Lee KW, Woon PS, Teo YY, Sim K. Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? *Neurosci Biobehav Rev* 2012; 36(1): 556-571.

72. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008; 32(4): 381-385.

73. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; 467(7317): 832-838.

74. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; 42(11): 937-948.

75. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 2008; 40(9): 1053-1055.

76. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D *et al.* Common variants conferring risk of schizophrenia. *Nature* 2009; 460(7256): 744-747.

77. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; 460(7256): 753-757.

78. Debnath M, Cannon DM, Venkatasubramanian G. Variation in the major histocompatibility complex [MHC] gene family in schizophrenia: associations and functional implications. *Prog Neuropsychopharmacol Biol Psychiatry* 2013; 42: 49-62.

79. Walters JT, Rujescu D, Franke B, Giegling I, Vasquez AA, Hargreaves A *et al.* The Role of the Major Histocompatibility Complex Region in Cognition and Brain Structure: A Schizophrenia GWAS Follow-Up. *Am J Psychiatry* 2013; 170(8): 877-885.

80. Flora A, Garcia JJ, Thaller C, Zoghbi HY. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc Natl Acad Sci U S A* 2007; 104(39): 15382-15387.

81. Brzozka MM, Radyushkin K, Wichert SP, Ehrenreich H, Rossner MJ. Cognitive and sensorimotor gating impairments in transgenic mice overexpressing the schizophrenia susceptibility gene Tcf4 in the brain. *Biol Psychiatry* 2010; 68(1): 33-40.

82. Steinberg S, de Jong S, Andreassen OA, Werge T, Borglum AD, Mors O *et al.* Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum Mol Genet* 2011; 20(20): 4076-4081.

83. Williams HJ, Norton N, Dwyer S, Moskvina V, Nikolov I, Carroll L *et al.* Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Mol Psychiatry* 2011; 16(4): 429-441.

84. Shi Y, Li Z, Xu Q, Wang T, Li T, Shen J *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat Genet* 2011; 43(12): 1224-1227.

85. Yue WH, Wang HF, Sun LD, Tang FL, Liu ZH, Zhang HX *et al.* Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat Genet* 2011; 43(12): 1228-1231.

86. Rietschel M, Mattheisen M, Degenhardt F, Kahn RS, Linszen DH, Os JV *et al.* Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe. *Mol Psychiatry* 2011.

87. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011; 43(10): 969-976.

88. Willemsen MH, Valles A, Kirkels LA, Mastebroek M, Olde Loohuis N, Kos A *et al.* Chromosome 1p21.3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability. *J Med Genet* 2011; 48(12): 810-818.

89. Kwon E, Wang W, Tsai LH. Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. *Mol Psychiatry* 2011.

90. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013; 45(10): 1150-1159.

91. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P *et al.* Large-scale copy number polymorphism in the human genome. *Science* 2004; 305(5683): 525-528.

92. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y *et al.* Detection of large-scale variation in the human genome. *Nat Genet* 2004; 36(9): 949-951.

93. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD *et al.* Global variation in copy number in the human genome. *Nature* 2006; 444(7118): 444-454.

94. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* 2010; 464(7289): 704-712.

95. Grayton HM, Fernandes C, Rujescu D, Collier DA. Copy number variations in neurodevelopmental disorders. *Prog Neurobiol* 2012; 99(1): 81-91.

96. Kirov G, Rees E, Walters JT, Escott-Price V, Georgieva L, Richards AL *et al.* The Penetrance of Copy Number Variations for Schizophrenia and Developmental Delay. *Biol Psychiatry* 2013.

97. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008; 320(5875): 539-543.

98. ISC. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; 455(7210): 237-241.

99. Xu B, Woodroffe A, Rodriguez-Murillo L, Roos JL, van Rensburg EJ, Abecasis GR *et al.* Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proc Natl Acad Sci U S A* 2009; 106(39): 16746-16751.

100. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 2012; 17(2): 142-153.

101. Kirov G, Gumus D, Chen W, Norton N, Georgieva L, Sari M *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* 2008; 17(3): 458-465.

102. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Sabatti C, Geurts van Kessel A *et al.* Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* 2008; 83(4): 504-510.

103. Rujescu D, Ingason A, Cichon S, Pietilainen OP, Barnes MR, Toulopoulou T *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* 2009; 18(5): 988-996.

104. Ikeda M, Aleksic B, Kirov G, Kinoshita Y, Yamanouchi Y, Kitajima T *et al.* Copy number variation in schizophrenia in the Japanese population. *Biol Psychiatry* 2010; 67(3): 283-286.

105. Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P *et al.* Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 2009; 18(8): 1497-1503.

106. Kirov G, Rujescu D, Ingason A, Collier DA, O'Donovan MC, Owen MJ. Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr Bull* 2009; 35(5): 851-854.

107. Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL *et al.* A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet* 2009; 5(2): e1000373.

108. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; 455(7210): 232-236.

109. Vacic V, McCarthy S, Malhotra D, Murray F, Chou HH, Peoples A *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* 2011; 471(7339): 499-503.

110. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 2009; 41(11): 1223-1227.

111. Ingason A, Rujescu D, Cichon S, Sigurdsson E, Sigmundsson T, Pietilainen OP *et al.* Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry* 2011; 16(1): 17-25.

112. Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 2010; 87(2): 229-236.

113. Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 2010; 87(5): 618-630.

114. Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z *et al.* Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PLoS One* 2013; 8(4): e59494.

115. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012; 337(6090): 64-69.

116. Shen YC, Liao DL, Chen JY, Wang YC, Lai IC, Liou YJ *et al.* Resequencing and association study of vesicular glutamate transporter 1 gene (VGLUT1) with schizophrenia. *Schizophr Res* 2009; 115(2-3): 254-260.

117. Shen YC, Liao DL, Lu CL, Chen JY, Liou YJ, Chen TT *et al.* Resequencing of the vesicular glutamate transporter 2 gene (VGLUT2) reveals some rare genetic variants that may increase the genetic burden in schizophrenia. *Schizophr Res* 2010; 121(1-3): 179-186.

118. Song W, Li W, Feng J, Heston LL, Scaringe WA, Sommer SS. Identification of high risk DISC1 structural variants with a 2% attributable risk for schizophrenia. *Biochem Biophys Res Commun* 2008; 367(3): 700-706.

119. Green EK, Grozeva D, Sims R, Raybould R, Forty L, Gordon-Smith K *et al.* DISC1 exon 11 rare variants found more commonly in schizoaffective spectrum cases than controls. *Am J Med Genet B Neuropsychiatr Genet* 2011; 156B(4): 490-492.

120. Moens LN, De Rijk P, Reumers J, Van den Bossche MJ, Glassee W, De Zutter S *et al.* Sequencing of DISC1 pathway genes reveals increased burden of rare missense variants in schizophrenia patients from a northern Swedish population. *PLoS One* 2011; 6(8): e23450.

121. Shen YC, Tsai HM, Cheng MC, Hsu SH, Chen SF, Chen CH. Genetic and functional analysis of the gene encoding neurogranin in schizophrenia. *Schizophr Res* 2012; 137(1-3): 7-13.

122. Smith RL, Knight D, Williams H, Dwyer S, Richards A, Kirov G *et al.* Analysis of neurogranin (NRGN) in schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2011; 156B(5): 532-535.

123. Dwyer S, Williams H, Holmans P, Moskvina V, Craddock N, Owen MJ *et al.* No evidence that rare coding variants in ZNF804A confer risk of schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2010; 153B(8): 1411-1416.

124. Ikeda M, Aleksic B, Kinoshita Y, Okochi T, Kawashima K, Kushima I *et al.* Genome-wide association study of schizophrenia in a Japanese population. *Biol Psychiatry* 2011; 69(5): 472-478.

125. Aleksic B, Kushima I, Hashimoto R, Ohi K, Ikeda M, Yoshimi A *et al.* Analysis of the VAV3 as Candidate Gene for Schizophrenia: Evidences From Voxel-Based Morphometry and Mutation Screening. *Schizophr Bull* 2012.

126. Kushima I, Nakamura Y, Aleksic B, Ikeda M, Ito Y, Shiino T *et al.* Resequencing and association analysis of the KALRN and EPHB1 genes and their contribution to schizophrenia susceptibility. *Schizophr Bull* 2012; 38(3): 552-560.

127. Mardis ER. Next-Generation Sequencing Platforms. *Annu Rev Anal Chem (Palo Alto Calif)* 2013

128. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011; 12(11): 745-755.

129. Lemaire M, Fremeaux-Bacchi V, Schaefer F, Choi M, Tang WH, Le Quintrec M *et al.* Recessive mutations in DGKE cause atypical hemolytic-uremic syndrome. *Nat Genet* 2013; 45(5): 531-536.

130. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010; 42(9): 790-793.

131. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet* 2012; 57(10): 621-632.

132. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 2012; 21(R1): R1-9.

133. Ezewudo M, Zwick ME. Evaluating rare variants in complex disorders using next-generation sequencing. *Curr Psychiatry Rep* 2013; 15(4): 349.

134. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* 2012; 44(6): 623-630.

135. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012; 336(6082): 740-743.

136. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; 467(7319): 1061-1073.

137. State MW, Levitt P. The conundrums of understanding genetic risks for autism spectrum disorders. *Nat Neurosci* 2011; 14(12): 1499-1506.

138. Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 2012; 44(12): 1365-1369.

139. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 2011; 43(9): 860-863.

140. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 2011; 43(9): 864-868.

141. Gratten J, Visscher PM, Mowry BJ, Wray NR. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet* 2013; 45(3): 234-238.

142. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR *et al.* Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* 2010; 87(3): 316-324.

143. Gauthier J, Champagne N, Lafreniere RG, Xiong L, Spiegelman D, Brustein E *et al.* De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. *Proc Natl Acad Sci U S A* 2010; 107(17): 7863-7868.

144. Tarabeux J, Champagne N, Brustein E, Hamdan FF, Gauthier J, Lapointe M *et al.* De novo truncating mutation in Kinesin 17 associated with schizophrenia. *Biol Psychiatry* 2010; 68(7): 649-656.

145. Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D, Maia JM *et al.* Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet* 2012; 91(2): 303-312.

146. Antonarakis SE, Chakravarti A, Cohen JC, Hardy J. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nat Rev Genet* 2010; 11(5): 380-384.

147. Timms AE, Dorschner MO, Wechsler J, Choi KY, Kirkwood R, Girirajan S *et al.* Support for the N -Methyl-D-Aspartate Receptor Hypofunction Hypothesis of Schizophrenia From Exome Sequencing in Multiplex Families. *JAMA Psychiatry* 2013: 1-9.

148. Sebat J, Levy DL, McCarthy SE. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet* 2009; 25(12): 528-535.

149. Tam GW, Redon R, Carter NP, Grant SG. The role of DNA copy number variation in schizophrenia. *Biol Psychiatry* 2009; 66(11): 1005-1012.

150. Muhleisen TW, Basmanav FB, Forstner AJ, Mattheisen M, Priebe L, Herms S *et al.* Resequencing and follow-up of neurexin 1 (NRXN1) in schizophrenia patients. *Schizophr Res* 2011; 127(1-3): 35-40.

151. Schmermund A, Mohlenkamp S, Stang A, Gronemeyer D, Seibel R, Hirche H *et al.* Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. *Am Heart J* 2002; 144(2): 212-218.

152. Borglum AD, Demontis D, Grove J, Pallesen J, Hollegaard MV, Pedersen CB *et al.* Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol Psychiatry* 2014; 19(3): 325-333.

153. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007; 35(6): 2013-2025.

154. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491(7422): 56-65.

155. Kral T, Clusmann H, Urbach J, Schramm J, Elger CE, Kurthen M *et al.* Preoperative evaluation for epilepsy surgery (Bonn Algorithm). *Zentralbl Neurochir* 2002; 63(3): 106-110.

156. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; 16(3): 1215.

157. Sela M, Anfinsen CB, Harrington WF. The correlation of ribonuclease activity with specific aspects of tertiary structure. *Biochim Biophys Acta* 1957; 26(3): 502-512.

158. Cardullo RA, Agrawal S, Flores C, Zamecnik PC, Wolf DE. Detection of nucleic acid hybridization by nonradiative fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A* 1988; 85(23): 8790-8794.

159. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001; 25(4): 402-408.

160. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977; 74(12): 5463-5467.

161. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* 1987; 155: 335-350.

162. Roux KH. Optimization and troubleshooting in PCR. *Cold Spring Harb Protoc* 2009; 2009(4): pdb ip66.

163. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res* 1995; 23(22): 4742-4743.

164. Tang K, Fu D, Kotter S, Cotter RJ, Cantor CR, Koster H. Matrix-assisted laser desorption/ionization mass spectrometry of immobilized duplex DNA probes. *Nucleic Acids Res* 1995; 23(16): 3126-3131.

165. Haff LA, Smirnov IP. Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res* 1997; 7(4): 378-388.

166. Cotter RJ. Time-of-flight mass spectrometry for the structural analysis of biological molecules. *Anal Chem* 1992; 64(21): 1027A-1039A.

167. Ragoussis J, Elvidge GP, Kaur K, Colella S. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet* 2006; 2(7): e100.

168. Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E *et al.* Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res* 2005; 33(6): e56.

169. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 2006; 7: 3.

170. Neumaier M, Braun A, Wagener C. Fundamentals of quality assessment of molecular amplification methods in clinical diagnostics. International Federation of Clinical Chemistry Scientific Division Committee on Molecular Biology Techniques. *Clin Chem* 1998; 44(1): 12-26.

171. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G *et al.* Impact of RNA degradation on gene expression profiling. *BMC Med Genomics* 2010; 3: 36.

172. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; 5(10): R80.

173. Ritchie ME, Dunning MJ, Smith ML, Shi W, Lynch AG. BeadArray expression analysis using bioconductor. *PLoS Comput Biol* 2011; 7(12): e1002276.

174. Sepp M, Kannike K, Eesmaa A, Urb M, Timmusk T. Functional diversity of human basic helix-loop-helix transcription factor TCF4 isoforms generated by alternative 5' exon usage and splicing. *PLoS One* 2011; 6(7): e22138.

175. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; 83(3): 311-321.

176. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; 5(2): e1000384.

177. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; 86(6): 832-838.

178. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011; 7(2): e1001289.

179. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; 89(1): 82-93.

180. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; 7(3): e1001322.

181. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R *et al.* 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics* 2012; 28(23): 3027-3033.

182. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* 1998; 23(6): 198-199.

183. Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, Hagemeier C *et al.* Binary specification of nonsense codons by splicing and cytoplasmic translation. *Embo J* 1998; 17(12): 3484-3494.

184. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. Nonsense-mediated decay approaches the clinic. *Nat Genet* 2004; 36(8): 801-808.

185. Isidor B, Lindenbaum P, Pichon O, Bezieau S, Dina C, Jacquemont S *et al.* Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. *Nat Genet* 2011; 43(4): 306-308.

186. Ozols J. Covalent structure of liver microsomal flavin-containing monooxygenase form 1. *J Biol Chem* 1990; 265(18): 10289-10299.

187. Fraaije MW, Kamerbeek NM, van Berkel WJ, Janssen DB. Identification of a Baeyer-Villiger monooxygenase sequence motif. *FEBS Lett* 2002; 518(1-3): 43-47.

188. Krueger SK, Williams DE. Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol Ther* 2005; 106(3): 357-387.

189. Polyakov AV, Shagina IA, Khlebnikova OV, Evgrafov OV. Mutation in the connexin 50 gene (GJA8) in a Russian family with zonular pulverulent cataract. *Clin Genet* 2001; 60(6): 476-478.

190. Poot M, van der Smagt JJ, Brilstra EH, Bourgeron T. Disentangling the myriad genomics of complex disorders, specifically focusing on autism, epilepsy, and schizophrenia. *Cytogenet Genome Res* 2011; 135(3-4): 228-240.

191. Hwang J, Kim YK. When a ribosome encounters a premature termination codon. *BMB Rep* 2013; 46(1): 9-16.

192. Bang ML, Owczarek S. A matter of balance: role of neurexin and neuroligin at the synapse. *Neurochem Res* 2013; 38(6): 1174-1189.

193. Sudhof TC. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 2008; 455(7215): 903-911.

194. Ullrich B, Ushkaryov YA, Sudhof TC. Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* 1995; 14(3): 497-507.

195. Lett TA, Tiwari AK, Meltzer HY, Lieberman JA, Potkin SG, Voineskos AN *et al.* The putative functional rs1045881 marker of neurexin-1 in schizophrenia and clozapine response. *Schizophr Res* 2011; 132(2-3): 121-124.

196. Reichelt AC, Rodgers RJ, Clapcote SJ. The role of neurexins in schizophrenia and autistic spectrum disorder. *Neuropharmacology* 2012; 62(3): 1519-1526.

197. Mei L, Xiong WC. Neuregulin 1 in neural development, synaptic plasticity and schizophrenia. *Nat Rev Neurosci* 2008; 9(6): 437-452.

198. Missler M, Zhang W, Rohlmann A, Kattenstroth G, Hammer RE, Gottmann K *et al.* Alpha-neurexins couple Ca2+ channels to synaptic vesicle exocytosis. *Nature* 2003; 423(6943): 939-948.

199. Dudanova I, Sedej S, Ahmad M, Masius H, Sargsyan V, Zhang W *et al.* Important contribution of alpha-neurexins to Ca2+-triggered exocytosis of secretory granules. *J Neurosci* 2006; 26(41): 10599-10613.

200. Zhang W, Rohlmann A, Sargsyan V, Aramuni G, Hammer RE, Sudhof TC *et al.* Extracellular domains of alpha-neurexins participate in regulating synaptic transmission by selectively affecting N- and P/Q-type Ca2+ channels. *J Neurosci* 2005; 25(17): 4330-4342.

201. Etherton MR, Blaiss CA, Powell CM, Sudhof TC. Mouse neurexin-1alpha deletion causes correlated electrophysiological and behavioral changes consistent with cognitive impairments. *Proc Natl Acad Sci U S A* 2009; 106(42): 17998-18003.

202. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010; 44: 293-308.

203. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 2011; 35(7): 606-619.

204. Bena F, Bruno DL, Eriksson M, van Ravenswaaij-Arts C, Stark Z, Dijkhuizen T *et al.* Molecular and clinical characterization of 25 individuals with exonic deletions of NRXN1 and comprehensive review of the literature. *Am J Med Genet B Neuropsychiatr Genet* 2013.

205. Ching MS, Shen Y, Tan WH, Jeste SS, Morrow EM, Chen X *et al.* Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. *Am J Med Genet B Neuropsychiatr Genet* 2010; 153B(4): 937-947.

206. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009; 459(7246): 569-573.

207. Kim HG, Kishikawa S, Higgins AW, Seong IS, Donovan DJ, Shen Y *et al.* Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* 2008; 82(1): 199-207.

208. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 2008; 82(2): 477-488.

209. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, Hill RS *et al.* Identifying autism loci and genes by tracing recent shared ancestry. *Science* 2008; 321(5886): 218-223.

210. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 2007; 39(3): 319-328.

211. Yan J, Noltner K, Feng J, Li W, Schroer R, Skinner C *et al.* Neurexin 1alpha structural variants associated with autism. *Neurosci Lett* 2008; 438(3): 368-370.

212. Feng J, Schroer R, Yan J, Song W, Yang C, Bockholt A *et al.* High frequency of neurexin 1beta signal peptide structural variants in patients with autism. *Neurosci Lett* 2006; 409(1): 10-13.

213. Shah AK, Tioleco NM, Nolan K, Locker J, Groh K, Villa C *et al.* Rare NRXN1 promoter variants in patients with schizophrenia. *Neurosci Lett* 2010; 475(2): 80-84.

214. Gauthier J, Siddiqui TJ, Huashan P, Yokomaku D, Hamdan FF, Champagne N *et al.* Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. *Hum Genet* 2011; 130(4): 563-573.

215. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012; 488(7412): 471-475.

216. Duong L, Klitten LL, Moller RS, Ingason A, Jakobsen KD, Skjodt C *et al.* Mutations in NRXN1 in a family multiply affected with brain disorders: NRXN1 mutations and brain disorders. *Am J Med Genet B Neuropsychiatr Genet* 2012; 159B(3): 354-358.

217. Rowen L, Young J, Birditt B, Kaur A, Madan A, Philipps DL *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* 2002; 79(4): 587-597.

218. Longmate JA, Larson GP, Krontiris TG, Sommer SS. Three ways of combining genotyping and resequencing in case-control association studies. *PLoS One* 2010; 5(12): e14318.

219. Jarinova O, Ekker M. Regulatory variations in the era of next-generation sequencing: implications for clinical molecular diagnostics. *Hum Mutat* 2012; 33(7): 1021-1030.

220. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 2011; 12(10): 683-691.

221. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM. The sounds of silence: synonymous mutations affect function. *Pharmacogenomics* 2007; 8(6): 527-532.

222. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 2012; 30(11): 1095-1106.

223. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489(7414): 57-74.

224. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010; 11(11): 773-785.

225. Crespi BJ, Crofts HJ. Association testing of copy number variants in schizophrenia and autism spectrum disorders. *J Neurodev Disord* 2012; 4(1): 15.

226. Grozeva D, Conrad DF, Barnes CP, Hurles M, Owen MJ, O'Donovan MC *et al.* Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. *Schizophr Res* 2012; 135(1-3): 1-7.

227. Tamminga CA, Stan AD, Wagner AD. The hippocampal formation in schizophrenia. *Am J Psychiatry* 2010; 167(10): 1178-1193.

228. Diogo D, Kurreeman F, Stahl EA, Liao KP, Gupta N, Greenberg JD *et al.* Rare, Low-Frequency, and Common Variants in the Protein-Coding Sequence of Biological Candidate Genes from GWASs Contribute to Risk of Rheumatoid Arthritis. *Am J Hum Genet* 2013; 92(1): 15-27.

229. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry* 2012; 17(5): 474-485.

230. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011; 43(11): 1066-1073.

231. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009; 324(5925): 387-389.

232. Harvard C, Strong E, Mercier E, Colnaghi R, Alcantara D, Chow E *et al.* Understanding the impact of 1q21.1 copy number variant. *Orphanet J Rare Dis* 2011; 6: 54.

233. Ye T, Lipska BK, Tao R, Hyde TM, Wang L, Li C *et al.* Analysis of copy number variations in brain DNA from patients with schizophrenia and other psychiatric disorders. *Biol Psychiatry* 2012; 72(8): 651-654.

234. Schweingruber C, Rufener SC, Zund D, Yamashita A, Muhlemann O. Nonsense-mediated mRNA decay - Mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochim Biophys Acta* 2013; 1829(6-7): 612-623.

235. Ji X, Li J, Zhu L, Cai J, Zhang J, Qu Y *et al.* CHD1L promotes tumor progression and predicts survival in colorectal carcinoma. *J Surg Res* 2013.

236. Flaus A, Martin DM, Barton GJ, Owen-Hughes T. Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res* 2006; 34(10): 2887-2905.

237. Ahel D, Horejsi Z, Wiechens N, Polo SE, Garcia-Wilson E, Ahel I *et al.* Poly(ADP-ribose)-dependent regulation of DNA repair by the chromatin remodeling enzyme ALC1. *Science* 2009; 325(5945): 1240-1243.

238. Pines A, Vrouwe MG, Marteijn JA, Typas D, Luijsterburg MS, Cansoy M *et al.* PARP1 promotes nucleotide excision repair through DDB2 stabilization and recruitment of ALC1. *J Cell Biol* 2012; 199(2): 235-249.

239. Gottschalk AJ, Trivedi RD, Conaway JW, Conaway RC. Activation of the SNF2 family ATPase ALC1 by poly(ADP-ribose) in a stable ALC1.PARP1.nucleosome intermediate. *J Biol Chem* 2012; 287(52): 43527-43532.

240. Gottschalk AJ, Timinszky G, Kong SE, Jin J, Cai Y, Swanson SK *et al.* Poly(ADP-ribosyl)ation directs recruitment and activation of an ATP-dependent chromatin remodeler. *Proc Natl Acad Sci U S A* 2009; 106(33): 13770-13774.

241. Li Y, Chen L, Chan TH, Liu M, Kong KL, Qiu JL *et al.* SPOCK1 is regulated by CHD1L and blocks apoptosis and promotes HCC cell invasiveness and metastasis in mice. *Gastroenterology* 2013; 144(1): 179-191 e174.

242. Chen L, Chan TH, Yuan YF, Hu L, Huang J, Ma S *et al.* CHD1L promotes hepatocellular carcinoma progression and metastasis in mice and is associated with these processes in human patients. *J Clin Invest* 2010; 120(4): 1178-1191.

243. Chen M, Huang JD, Hu L, Zheng BJ, Chen L, Tsang SL *et al.* Transgenic CHD1L expression in mouse induces spontaneous tumors. *PLoS One* 2009; 4(8): e6727.

244. Ma NF, Hu L, Fung JM, Xie D, Zheng BJ, Chen L *et al.* Isolation and characterization of a novel oncogene, amplified in liver cancer 1, within a commonly amplified region at 1q21 in hepatocellular carcinoma. *Hepatology* 2008; 47(2): 503-510.

245. Chan TH, Chen L, Liu M, Hu L, Zheng BJ, Poon VK *et al.* Translationally controlled tumor protein induces mitotic defects and chromosome missegregation in hepatocellular carcinoma development. *Hepatology* 2012; 55(2): 491-505.

246. Hyeon J, Ahn S, Park CK. CHD1L Is a Marker for Poor Prognosis of Hepatocellular Carcinoma after Surgical Resection. *Korean J Pathol* 2013; 47(1): 9-15.

247. Brockschmidt A, Chung B, Weber S, Fischer DC, Kolatsi-Joannou M, Christ L *et al.* CHD1L: a new candidate gene for congenital anomalies of the kidneys and urinary tract (CAKUT). *Nephrol Dial Transplant* 2012; 27(6): 2355-2364.

248. He WP, Zhou J, Cai MY, Xiao XS, Liao YJ, Kung HF *et al.* CHD1L protein is overexpressed in human ovarian carcinomas and is a novel predictive biomarker for patients survival. *BMC Cancer* 2012; 12: 437.

249. Preti A, Wilson DR. Schizophrenia, cancer and obstetric complications in an evolutionary perspective-an empirically based hypothesis. *Psychiatry Investig* 2011; 8(2): 77-88.

250. Wang Y, He G, He L, McGrath J. Do shared mechanisms underlying cell cycle regulation and synaptic plasticity underlie the reduced incidence of cancer in schizophrenia? *Schizophr Res* 2011; 130(1-3): 282-284.

251. Cui DH, Jiang KD, Jiang SD, Xu YF, Yao H. The tumor suppressor adenomatous polyposis coli gene is associated with susceptibility to schizophrenia. *Mol Psychiatry* 2005; 10(7): 669-677.

252. Kalkman HO. The role of the phosphatidylinositide 3-kinase-protein kinase B pathway in schizophrenia. *Pharmacol Ther* 2006; 110(1): 117-134.

253. Matigian N, Abrahamsen G, Sutharsan R, Cook AL, Vitale AM, Nouwens A *et al.* Disease-specific, neurosphere-derived cells as models for brain disorders. *Dis Model Mech* 2010; 3(11-12): 785-798.

254. Fan Y, Abrahamsen G, McGrath JJ, Mackay-Sim A. Altered cell cycle dynamics in schizophrenia. *Biol Psychiatry* 2012; 71(2): 129-135.

255. Kerns D, Vong GS, Barley K, Dracheva S, Katsel P, Casaccia P *et al.* Gene expression abnormalities and oligodendrocyte deficits in the internal capsule in schizophrenia. *Schizophr Res* 2010; 120(1-3): 150-158.

256. Katsel P, Davis KL, Li C, Tan W, Greenstein E, Kleiner Hoffman LB *et al.* Abnormal indices of cell cycle activity in schizophrenia and their potential association with oligodendrocytes. *Neuropsychopharmacology* 2008; 33(12): 2993-3009.

257. Benes FM, Lim B, Subburaju S. Site-specific regulation of cell cycle and DNA repair in post-mitotic GABA cells in schizophrenic versus bipolars. *Proc Natl Acad Sci U S A* 2009; 106(28): 11731-11736.

258. Lange C, Huttner WB, Calegari F. Cdk4/cyclinD1 overexpression in neural stem cells shortens G1, delays neurogenesis, and promotes the generation and expansion of basal progenitors. *Cell Stem Cell* 2009; 5(3): 320-331.

259. Frank CL, Tsai LH. Alternative functions of core cell cycle regulators in neuronal migration, neuronal maturation, and synaptic plasticity. *Neuron* 2009; 62(3): 312-326.

260. Kurian SM, Le-Niculescu H, Patel SD, Bertram D, Davis J, Dike C *et al.* Identification of blood biomarkers for psychosis using convergent functional genomics. *Mol Psychiatry* 2011; 16(1): 37-58.

261. Tolias KF, Duman JG, Um K. Control of synapse development and plasticity by Rho GTPase regulatory proteins. *Prog Neurobiol* 2011; 94(2): 133-148.

262. Kneussel M, Engelkamp D, Betz H. Distribution of transcripts for the brain-specific GDP/GTP exchange factor collybistin in the developing mouse brain. *Eur J Neurosci* 2001; 13(3): 487-492.

263. Reid T, Bathoorn A, Ahmadian MR, Collard JG. Identification and characterization of hPEM-2, a guanine nucleotide exchange factor specific for Cdc42. *J Biol Chem* 1999; 274(47): 33587-33593.

264. Gonzalez-Billault C, Munoz-Llancao P, Henriquez DR, Wojnacki J, Conde C, Caceres A. The role of small GTPases in neuronal morphogenesis and polarity. *Cytoskeleton (Hoboken)* 2012; 69(7): 464-485.

265. Scott EK, Reuter JE, Luo L. Small GTPase Cdc42 is required for multiple aspects of dendritic morphogenesis. *J Neurosci* 2003; 23(8): 3118-3123.

266. Vadodaria KC, Brakebusch C, Suter U, Jessberger S. Stage-specific functions of the small Rho GTPases Cdc42 and Rac1 for adult hippocampal neurogenesis. *J Neurosci* 2013; 33(3): 1179-1189.

267. Chen C, Wirth A, Ponimaskin E. Cdc42: an important regulator of neuronal morphology. *Int J Biochem Cell Biol* 2012; 44(3): 447-451.

268. Hill JJ, Hashimoto T, Lewis DA. Molecular mechanisms contributing to dendritic spine alterations in the prefrontal cortex of subjects with schizophrenia. *Mol Psychiatry* 2006; 11(6): 557-566.

269. Ide M, Lewis DA. Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. *Biol Psychiatry* 2010; 68(1): 25-32.

270. Tyagarajan SK, Ghosh H, Harvey K, Fritschy JM. Collybistin splice variants differentially interact with gephyrin and Cdc42 to regulate gephyrin clustering at GABAergic synapses. *J Cell Sci* 2011; 124(Pt 16): 2786-2796.

271. Lionel AC, Vaags AK, Sato D, Gazzellone MJ, Mitchell EB, Chen HY *et al.* Rare exonic deletions implicate the synaptic organizer Gephyrin (GPHN) in risk for autism, schizophrenia and seizures. *Hum Mol Genet* 2013; 22(10): 2055-2066.

272. Chen L, Hu L, Chan TH, Tsao GS, Xie D, Huo KK *et al.* Chromodomain helicase/adenosine triphosphatase DNA binding protein 1-like (CHD1l) gene suppresses the nucleus-to-mitochondria translocation of nur77 to sustain hepatocellular carcinoma cell survival. *Hepatology* 2009; 50(1): 122-129.

273. Maxwell MA, Muscat GE. The NR4A subgroup: immediate early response genes with pleiotropic physiological roles. *Nucl Recept Signal* 2006; 4: e002.

274. Levesque D, Rouillard C. Nur77 and retinoid X receptors: crucial factors in dopamine-related neuroadaptation. *Trends Neurosci* 2007; 30(1): 22-30.

275. Xing G, Zhang L, Russell S, Post R. Reduction of dopamine-related transcription factors Nurr1 and NGFI-B in the prefrontal cortex in schizophrenia and bipolar disorders. *Schizophr Res* 2006; 84(1): 36-56.

276. Jiang P, Zhang WY, Li HD, Cai HL, Xue Y. Repeated haloperidol administration has no effect on vitamin D signaling but increase retinoid X receptors and Nur77 expression in rat prefrontal cortex. *Cell Mol Neurobiol* 2013; 33(3): 309-312.

277. Maheux J, Ethier I, Rouillard C, Levesque D. Induction patterns of transcription factors of the nur family (nurr1, nur77, and nor-1) by typical and atypical antipsychotics in the mouse brain: implication for their mechanism of action. *J Pharmacol Exp Ther* 2005; 313(1): 460-473.

278. Eells JB, Wilcots J, Sisk S, Guo-Ross SX. NR4A gene expression is dynamically regulated in the ventral tegmental area dopamine neurons and is related to expression of dopamine neurotransmission genes. *J Mol Neurosci* 2012; 46(3): 545-553.

279. Gilbert F, Morissette M, St-Hilaire M, Paquet B, Rouillard C, Di Paolo T *et al.* Nur77 gene knockout alters dopamine neuron biochemical activity and dopamine turnover. *Biol Psychiatry* 2006; 60(6): 538-547.

280. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012; 485(7397): 242-245.

281. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 2012; 338(6114): 1619-1622.

282. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; 485(7397): 246-250.

283. Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* 2013; 92(2): 221-237.

284. Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 2008; 40(12): 1466-1471.

285. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 2008; 359(16): 1685-1699.

286. Cashman JR. Structural and catalytic properties of the mammalian flavin-containing monooxygenase. *Chem Res Toxicol* 1995; 8(2): 166-181.

287. Zhang J, Cashman JR. Quantitative analysis of FMO gene mRNA levels in human tissues. *Drug Metab Dispos* 2006; 34(1): 19-26.

288. Janmohamed A, Hernandez D, Phillips IR, Shephard EA. Cell-, tissue-, sex- and developmental stage-specific expression of mouse flavin-containing monooxygenases (Fmos). *Biochem Pharmacol* 2004; 68(1): 73-83.

289. Gagliardi S, Ogliari P, Davin A, Corato M, Cova E, Abel K *et al.* Flavin-containing monooxygenase mRNA levels are up-regulated in als brain areas in SOD1-mutant mice. *Neurotox Res* 2011; 20(2): 150-158.

290. Gagliardi S, Abel K, Bianchi M, Milani P, Bernuzzi S, Corato M *et al.* Regulation of FMO and PON detoxication systems in ALS human tissues. *Neurotox Res* 2013
23(4): 370-377.

291. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 2010; 38(3): e17.

292. Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavare S, Ritchie ME. Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* 2008; 9: 85.

293. Wu JQ, Kosten TR, Zhang XY. Free radicals, antioxidant defense systems, and schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry* 2013.

294. Hibar DP, Stein JL, Ryles AB, Kohannim O, Jahanshad N, Medland SE *et al.* Genome-wide association identifies genetic variants associated with lentiform nucleus volume in N = 1345 young and elderly subjects. *Brain Imaging Behav* 2013; 7(2): 102-115.

295. Zhang J, Cerny MA, Lawson M, Mosadeghi R, Cashman JR. Functional activity of the mouse flavin-containing monooxygenase forms 1, 3, and 5. *J Biochem Mol Toxicol* 2007; 21(4): 206-215.

296. Sohl G, Maxeiner S, Willecke K. Expression and functions of neuronal gap junctions. *Nat Rev Neurosci* 2005; 6(3): 191-200.

297. Wei CJ, Xu X, Lo CW. Connexins and cell signaling in development and disease. *Annu Rev Cell Dev Biol* 2004; 20: 811-838.

298. Cai G, Wu X, Li T, Collier DA, Liu X, Feng B *et al.* [Linkage analysis of susceptibility genes for familial schizophrenia on chromosome 1 in Chinese population]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 2002; 19(6): 491-494.

299. Ekelund J, Hovatta I, Parker A, Paunio T, Varilo T, Martin R *et al.* Chromosome 1 loci in Finnish schizophrenia families. *Hum Mol Genet* 2001; 10(15): 1611-1617.

300. Hwu HG, Liu CM, Fann CS, Ou-Yang WC, Lee SF. Linkage of schizophrenia with chromosome 1q loci in Taiwanese families. *Mol Psychiatry* 2003; 8(4): 445-452.

301. Brzustowicz LM, Hodgkinson KA, Chow EW, Honer WG, Bassett AS. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. *Science* 2000; 288(5466): 678-682.

302. Adachi S, Jigami T, Yasui T, Nakano T, Ohwada S, Omori Y *et al.* Role of a BCL9-related beta-catenin-binding protein, B9L, in tumorigenesis induced by aberrant activation of Wnt signaling. *Cancer Res* 2004; 64(23): 8496-8501.

303. Kramps T, Peter O, Brunner E, Nellen D, Froesch B, Chatterjee S *et al.* Wnt/wingless signaling requires BCL9/legless-mediated recruitment of pygopus to the nuclear beta-catenin-TCF complex. *Cell* 2002; 109(1): 47-60.

304. Stadeli R, Hoffmans R, Basler K. Transcription under the control of nuclear Arm/beta-catenin. *Curr Biol* 2006; 16(10): R378-385.

305. Inestrosa NC, Montecinos-Oliva C, Fuenzalida M. Wnt signaling: role in Alzheimer disease and schizophrenia. *J Neuroimmune Pharmacol* 2012; 7(4): 788-807.

306. Maguschak KA, Ressler KJ. A role for WNT/beta-catenin signaling in the neural mechanisms of behavior. *J Neuroimmune Pharmacol* 2012; 7(4): 763-773.

307. Meffre D, Grenier J, Bernard S, Courtin F, Dudev T, Shackleford G *et al.* Wnt and lithium: a common destiny in the therapy of nervous system pathologies? *Cell Mol Life Sci* 2013.

308. Singh KK. An emerging role for Wnt and GSK3 signaling pathways in schizophrenia. *Clin Genet* 2013; 83(6): 511-517.

309. Hansen T, Ingason A, Djurovic S, Melle I, Fenger M, Gustafsson O *et al.* At-risk variant in TCF7L2 for type II diabetes increases risk of schizophrenia. *Biol Psychiatry* 2011; 70(1): 59-63.

310. Kishimoto M, Ujike H, Okahisa Y, Kotaka T, Takaki M, Kodama M *et al.* The Frizzled 3 gene is associated with methamphetamine psychosis in the Japanese population. *Behav Brain Funct* 2008; 4: 37.

311. Proitsi P, Li T, Hamilton G, Di Forti M, Collier D, Killick R *et al.* Positional pathway screen of wnt signaling genes in schizophrenia: association with DKK4. *Biol Psychiatry* 2008; 63(1): 13-16.

312. Yang J, Si T, Ling Y, Ruan Y, Han Y, Wang X *et al.* Association study of the human FZD3 locus with schizophrenia. *Biol Psychiatry* 2003; 54(11): 1298-1301.

313. Li J, Zhou G, Ji W, Feng G, Zhao Q, Liu J *et al.* Common variants in the BCL9 gene conferring risk of schizophrenia. *Arch Gen Psychiatry* 2011; 68(3): 232-240.

314. Li M, Zhang H, Luo XJ, Gao L, Qi XB, Gourraud PA *et al.* Meta-Analysis Indicates That the European GWAS-Identified Risk SNP rs1344706 within ZNF804A Is Not Associated with Schizophrenia in Han Chinese Population. *PLoS One* 2013; 8(6): e65780.

315. Xu C, Aragam N, Li X, Villla EC, Wang L, Briones D *et al.* BCL9 and C9orf5 are associated with negative symptoms in schizophrenia: meta-analysis of two genome-wide association studies. *PLoS One* 2013; 8(1): e51674.

316. Carroll LS, Williams HJ, Walters J, Kirov G, O'Donovan MC, Owen MJ. Mutation screening of the 3q29 microdeletion syndrome candidate genes DLG1 and PAK2 in schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2011; 156B(7): 844-849.

317. Ravanpay AC, Olson JM. E protein dosage influences brain development more than family member identity. *J Neurosci Res* 2008; 86(7): 1472-1481.

318. Persson P, Jogi A, Grynfeld A, Pahlman S, Axelson H. HASH-1 and E2-2 are expressed in human neuroblastoma cells and form a functional complex. *Biochem Biophys Res Commun* 2000; 274(1): 22-31.

319. Fu H, Cai J, Clevers H, Fast E, Gray S, Greenberg R *et al.* A genome-wide screen for spatially restricted expression patterns identifies transcription factors that regulate glial development. *J Neurosci* 2009; 29(36): 11399-11408.

320. Navarrete K, Pedroso I, De Jong S, Stefansson H, Steinberg S, Stefansson K *et al.* TCF4 (e2-2; ITF2): a schizophrenia-associated gene with pleiotropic effects on human disease. *Am J Med Genet B Neuropsychiatr Genet* 2012; 162(1): 1-16.

321. Wirgenes KV, Sonderby IE, Haukvik UK, Mattingsdal M, Tesli M, Athanasiu L *et al.* TCF4 sequence variants and mRNA levels are associated with neurodevelopmental characteristics in psychotic disorders. *Transl Psychiatry* 2012; 2: e112.

322. Kwon E, Wang W, Tsai LH. Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. *Mol Psychiatry* 2013; 18(1): 11-12.

323. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* 2011 108(29): 11983-11988.

324. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; 44(3): 243-246.

325. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012; 337(6090): 100-104.

326. Hu X, Zhang B, Liu W, Paciga S, He W, Lanz TA *et al.* A survey of rare coding variants in candidate genes in schizophrenia by deep sequencing. *Mol Psychiatry* 2013.

327. Forrest M, Chapman RM, Doyle AM, Tinsley CL, Waite A, Blake DJ. Functional analysis of TCF4 missense mutations that cause Pitt-Hopkins syndrome. *Hum Mutat* 2012.

328. Sepp M, Pruunsild P, Timmusk T. Pitt-Hopkins syndrome-associated mutations in TCF4 lead to variable impairment of the transcription factor function ranging from hypomorphic to dominant-negative effects. *Hum Mol Genet* 2012; 21(13): 2873-2888.

329. Blake DJ, Forrest M, Chapman RM, Tinsley CL, O'Donovan MC, Owen MJ. TCF4, schizophrenia, and Pitt-Hopkins Syndrome. *Schizophr Bull* 2010; 36(3): 443-447.

330. Friedrich FW, Dilanian G, Khattar P, Juhr D, Gueneau L, Charron P *et al.* A novel genetic variant in the transcription factor Islet-1 exerts gain of function on myocyte enhancer factor 2C promoter activity. *Eur J Heart Fail* 2013; 15(3): 267-276.

331. Perrotta S, Stiehl DP, Punzo F, Scianguetta S, Borriello A, Bencivenga D *et al.* Congenital erythrocytosis associated with gain-of-function HIF2A gene mutations and erythropoietin levels in the normal range. *Haematologica* 2013; 98(10): 1624-1632.

332. Yagnik G, Ghuman A, Kim S, Stevens CG, Kimonis V, Stoler J *et al.* ALX4 gain-of-function mutations in nonsyndromic craniosynostosis. *Hum Mutat* 2012; 33(12): 1626-1629.

333. Bonnefond A, Clement N, Fawcett K, Yengo L, Vaillant E, Guillaume JL *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet* 2012; 44(3): 297-301.

334. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* 2010; 42(8): 684-687.

335. Dow DJ, Huxley-Jones J, Hall JM, Francks C, Maycox PR, Kew JN *et al.* ADAMTSL3 as a candidate gene for schizophrenia: gene sequencing and ultra-high density association analysis by imputation. *Schizophr Res* 2011; 127(1-3): 28-34

8.LIST OF PUBLICATIONS

Forstner AJ*, Basmanav FB*, Mattheisen M, GROUP Investigators, Bohmer AC, Hollegaard MV, Janson E, Strengman E, Priebe L, Degenhardt F, Hoffmann P, Herms S, Maier W, Mossner R, Rujescu D, Ophoff RA, Moebus S, Mortensen PB, Borglum AD, Hougaard DM, Frank J, Witt SH, Rietschel M, Zimmer A, Nothen MM*, Miro X*, Cichon S*, ***Investigation of the involvement of MIR185 and its target genes in the development of schizophrenia***, J Psychiatry Neurosci. 2014 Jun 17;39(4):130189.]

Basmanav FB*, Oprisoreanu AM*, Pasternack SM*, Thiele H, Fritz G, Wenzel J, Größer L, Wehner M, Wolf S, Fagerberg C, Bygum A, Altmüller J, Rütten A, Parmentier L, El Shabrawi-Caelen L, Hafner C, Nürnberg P, Kruse R, Schoch S, Hanneken S, Betz RC, ***Mutations in POGLUT1, encoding protein O-glucosyltransferase 1, cause autosomal-dominant Dowling-Degos disease***, Am J Hum Genet. 2014 Jan 2;94(1):135-43.

Nenadic I, Maitra R, Basmanav FB, Schultz CC, Lorenz C, Schachtzabel C, Smesny S, Nöthen MM, Cichon S, Reichenbach JR, Sauer H, Schlösser RG, Gaser C, ***ZNF804A genetic variation (rs1344706) affects brain grey but not white matter in schizophrenia and healthy subjects*** [accepted for publication in Psychological Medicine]

Schultz CC*, Nenadic I*, Riley B, Vladimirov VI, Wagner G, Koch K, Schachtzabel C, Mühleisen TW, Basmanav B, Nöthen MM, Deufel T, Kiehntopf M, Rietschel M, Reichenbach JR, Cichon S, Schlösser RG*, Sauer H*, ***ZNF804A and Cortical Structure in Schizophrenia: In Vivo and Postmortem Studies***, Schizophr Bull. 2014 May;40(3):532-41.

Rietschel M*, Mattheisen M*, Degenhardt F; GROUP Investigators; Genetic Risk and Outcome in Psychosis (GROUP Investigators), Kahn RS, Linszen DH, Os JV, Wiersma D, Bruggeman R, Cahn W, de Haan L, Krabbendam L, Myin-Germeys I, Mühleisen TW, Kirsch P, Esslinger C, Herms S, Demontis D, Steffens M, Strohmaier J, Haenisch B, Breuer R, Czerski PM, Giegling I, Strengman E, Schmael C, Mors O, Mortensen PB, Hougaard DM, Orntoft T, Kapelski P, Priebe L, Basmanav FB, Forstner AJ, Hoffmann P, Meier S, Nikitopoulos J, Moebus S, Alexander M, Mössner R, Wichmann HE, Schreiber S, Rivandeneira F, Hofman A, Uitterlinden AG, Wienker TF, Schumacher J, Hauser J, Maier W, Cantor RM, Erk S, Schulze TG; SGENE-plus Consortium; (Only those persons responsible for the samples of Replication 2 are listed), Stefansson H, Steinberg S, Gustafsson O, Sigurdsson E, Petursson H, Kong A, Stefansson K, Pietiläinen OP, Tuulio-Henriksson A, Paunio T, Lonnqvist J, Suvisaari J, Peltonen L, Ruggeri M, Tosato S, Walshe M, Murray R, Collier DA, Clair DS, Hansen T, Ingason A, Jakobsen KD, Duong L, Werge T, Melle I, Andreassen OA, Djurovic S, Bitter I, Réthelyi JM, Abramova L, Kaleda V, Golimbet V, Jönsson EG, Terenius L, Agartz I, Winkel RV, Kenis G, Hert MD, Veldink J, Wiuf C, Didriksen M, Craddock N, Owen MJ, O'Donovan MC, Børglum AD, Rujescu D, Walter H, Meyer-Lindenberg A, Nöthen MM*, Ophoff RA*, Cichon S*, ***Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe***, Mol Psychiatry. 2012 Sep;17(9):906-17.

Mühleisen TW, Basmanav FB, Forstner AJ, Mattheisen M, Priebe L, Herms S, Breuer R, Moebus S, Nenadic I, Sauer H, Mössner R, Maier W, Rujescu D, Ludwig M, Rietschel M, Nöthen MM, Cichon S, ***Resequencing and follow-up of neurexin 1 (NRXN1) in schizophrenia patients***, Schizophr Res. 2011 Apr;127(1-3):35-40.

Zorlutuna P, Yilgor P, Basmanav FB, Hasirci V, ***Biomaterials and tissue engineering research in Turkey: The METU Biomat Center experience***, Biotechnol J. 2009 Jul;4(7):965-80.

Basmanav FB, Kose GT, Hasirci V, ***Sequential Growth Factor Delivery From Complexed Microspheres for Bone Tissue Engineering***, Biomaterials, 2008, 29(31):4195-204.

Hasirci V, Vrana E, Zorlutuna P, Ndreu A, Yilgor P, Basmanav B, Aydin E, ***Nanobiomaterials; A Review of the Existing Science and Technology, and New Approaches***, J Biomater Sci Polym Ed. 2006;17(11):1241-68.

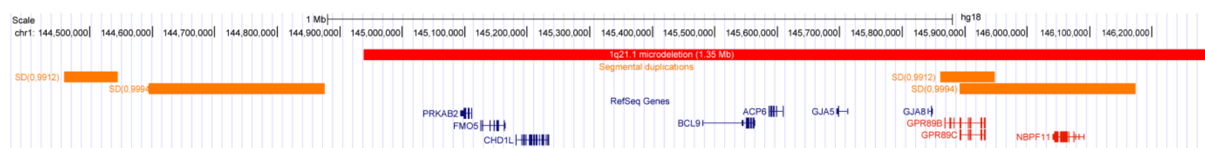
* Authors contributed equally to this work.

9. ATTACHMENTS

I. List of genes that were investigated in the current thesis. RefSeq IDs of the targeted transcripts and Uniprot IDs for their respective proteins are given for the *NRXN1* gene, 7 genes in the 1q21.1 microdeletion region (*PRKAB2*, *FMO5*, *CHD1L*; *BCL9*, *ACP6*, *GJA5*, and *GJA8*) and the *TCF4* gene.

Gene	RefSeq ID	Uniprot ID	Length (AA)
NRXN1α2	NM_001135659	Q9ULB1-3 (NRX1A_HUMAN)	1547
NRXN1β	NM_138735	P58400 (NRX1B_HUMAN)	442
PRKAB2	NM_005399.3	O43741 (AAKB2_HUMAN)	272
FMO5	NM_001461.2	P49326 (FMO5_HUMAN)	533
CHD1L	NM_004284.3	Q86WJ1 (CHD1L_HUMAN)	897
BCL9	NM_004326.2	O00512 (BCL9_HUMAN)	1426
ACP6	NM_016361.3	Q9NPH0 (PPA6_HUMAN)	428
GJA5	NM_005266.5	P36382 (CXA5_HUMAN)	358
GJA8	NM_005267.4	P48165 (CXA8_HUMAN)	433
TCF4	NM_001083962.1	P15884-3 (ITF2_HUMAN)	671

II. Genomic view of the 1q21.1 microdeletion region. Overview of the short form of the 1q21.1 microdeletion region (~1.35 Mb) was obtained from the UCSC Genome Browser. The microdeletion spans ten protein coding RefSeq genes. Three of these genes, namely *GPR89B*; *GPR89C* and *NBPF11*, are located in the flanking segmental duplications (SD).



III. Primers used for resequencing of the targeted genes. 25 primer pairs were designed targeting 24 coding exons of the *NRXN1* gene. 67 primer pairs were designed targeting in total 57 coding exons of the *PRKAB2*, *FMO5*, *CHD1L*, *BCL9*, *ACP6*, *GJA5* and *GJA8* genes located in the 1q21.1 microdeletion region. 18 primer pairs were designed targeting 18 protein coding exons of the *TCF4* gene. A single primer pair spanning exons 4-10 was designed for *CHD1L* cDNA sequencing.

NRXN1	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 2	GGGAGTAGTGTTGGTAGAAAGGTC	GAACTAGATAAAGGAGGGCACATC	1175
Exon 3	GTTGATTGCCTTGCTTTGAC	GTTCTTGGGGCTTAAAACAGTG	266
Exon 4	GGGTTGAGAATGTATGTGTTCTG	ATGATGTGTTTGGTGCCTTG	294
Exon 5	GCAGCCATATAATTTGCAAGC	CTGTCACTGAACATTTTTGTCACC	228
Exon 6	CTTATTTGCCACGAACTGTG	GCTGCAGAACAAGAGAAGTTAG	196
Exon 7	GGAGCCCTGTATCATGTTGTTAG	CTATTGTTTCTGTTAAGACCTGCTG	475
Exon 8	AAGAGTACCAACCGCAGTTC	GTTTCATTCTACTAACCGTTACCTG	282
Exon 9	GTGCTTTCATGCTGGATCTG	CAGGCATATCCCAGGATTACAG	303
Exon 10	TGAAACAGAAGCAATATCAGGC	TCGTTGAAAGTTACATGAGCTG	672
Exon 11	AGTGCAGGTTTGAGGTCAAC	CATCAAGCAGACGAAAGAGAG	542
Exon 12	CAAACAGGCTACTTGATCAATG	TACTGTGACTTCTCTTTCCAGC	467
Exon 13	AGTGGGAAAGTCTTCAGCTGTAC	TGGCCATATTTGCATGTGTC	332
Exon 14	ATTTCAAGTTGTGTACCAGCC	CCCTTGTCTCTTTCTCTCACTC	310
Exon 15	GAGCCAGTATGTTCTTCTTAGTG	GAGTACCAAGGAGGGATTTG	593
Exon 16	CTTTGCAGAAGGTACAAACACAC	ACTTCTAGTAAGGATGGAACCACC	405
Exon 17	GGGTATTTGACCAGTTATCAG	TTTCTCCATCCTACTGTGTG	351
Exon 18	TTCTATGGGTTATGACAGTTCG	GCAACTACTACATGTGATCTAGCTC	332
Exon1(β)	GTCTTCAAAGAGATAAGTGGCTCG	CCGCACACAAAGCTAAATGG	886
Exon 19	GAAGGAAGCATCCAAGAAGC	TGGTCCACTTTCTTGAGCAC	433
Exon 20	ACAATTTGGTGAAACGGATG	GGAGGAAAGCAGCTGTTGTC	376
Exon 21	GTTAATGCTTGCTGTGAGTACC	ATCATGTGGCAACTGTATAGCTC	481
Exon 22	AGACGCATATGCAGAAAAGC	ATGCATTTCATGGACACCTC	563
Exon 23	AGGAGGGTAGCCTTCTATACATG	GGATCTAAAATGTAAGCTCTGTGTC	345
Exon 24	TTTCCTTCCTGATTGCATTC	TTAGTAACCATGAGGCAAAGG	598

PRKAB2	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 8	ATCAGCCTTCCAGTCTCAGG	TGACTGTGAAGGACTGATGACTG	252
Exon 7	TAGCAGAGAAAGTAACCACTGAG	CTGAGGAAGAGTAAGCTGTTTG	240
Exon 6	TCAACACACCTCTGTAGTAATCC	AGCTGCCATGTATTGATAGC	318
Exon 5	AACTGGCAGGATTAAGATGG	CACTGGAGGATTTCTCTCTTG	666
Exon 4	TCTCTTCCATCCAACCTTCC	CAGGGGATACAATGGCTTAGG	398
Exon 3	GTCCAGCTGGGTCTCTCTTC	GCCTAGCACCCTTCTCTATCCTA	341
Exon 2	CGGCACTTCAAGAGGAATTATC	GATAGCGGGTTTCCTGAGC	555

III. Primers used for the resequencing of targeted genes -cont'd.

FMO5	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 9	AGAGTCAATCTCGTCAGATTCTG	GTCTGATATTGAGCCAATAGCAG	524
Exon 8	CCTTAAGCTAACATGAGGCAG	CCTGTATTTCTTTGAATTCTCC	227
Exon 7	AAACAAGTACCTAAACAGAGG	ATTCTGTTTACCTCCACC	516
Exon 6	TTAATGTATGGGTAGAGGTGGC	GCCTGAAATGATTCTTGAAGAA	389
Exon 5	AGGTATCCCTATTCTCTGTTGG	ACAAATGACTTGACAAATGATATC	308
Exon 4	CCAAAGTAGGAATTACTCAGACC	GATAGTGCCTGTGATATCAGC	350
Exon 3	CCAATCTTTAGAAGTAAACATAGG	CTTTTTCAGATGTCTATTGGC	363
Exon 2	GCCACCTGACACTGTTAAGAT	AAATGTTAACCGATGTGTGC	307

CHD1L	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 1	AAGCACGACTGGTGAGAACTG	CGCTGATCTCACCACGTTTC	644
Exon 2	CAAACTCTTAGAACATTGCCTC	AGGGACATTTGTGAAACTGAG	300
Exon 3	TGAGGCTTACATGATTTCGC	TTACTACAAACGATTGCTGACC	512
Exon 4	TTCCAGATATGCACATCCTG	CTGATGAAGGTAGACTCCCTG	313
Exon 5	AATCACGTCTGTTACTGGAG	ACCAGACATGCCAGGAGAGG	421
Exon 6	ACATGTAGCCTCTCAGGAAG	GTGAAGCTGTACAAGTGTGG	416
Exon 7	GTGTGGGTAAAGGTTGTGAT	TTCTACCTATAACCATTATGCATG	325
Exon 8	AGCTGTTATGCTGATCAATG	TAAGATGCTCTGCCACAGTC	623
Exon 9	TTGAAATGATTAAGGTTACTTG	GAAGAATTTTGGTATGATTTG	266
Exon 10	TGATGAGCTTTTGGCTTCTC	GGAATTGGCCTGAACAACTC	280
Exon 11	ATTGTTTGGTAAATTTTGCC	TTCATGGACCTATCTGGACT	225
Exon 12	AAATTGGCCTTTGGTTTTGG	AAAGTGTACTAACCGTTGGGATG	282
Exon 13	TGACCCACTCTAGCCTTGTG	GGACTGTCACCAGGACTATATTC	284
Exon 14	CTCAGGACCAAATGAACCAATC	AGCACTGTGGCAGGTGTGAG	447
Exon 15	TGTGCCTGGTGTCGTTAATC	CAGAGAGCAGCATCCAAAAC	351
Exon 16	ATAGGCTGTTCTCCTGTCCAC	GTCATCATTACCAACAGCAGG	567
Exon 17	TTCCGTACAGTGTGTGTTAGG	TGATATTCAACTGTGGTCTCAC	323
Exon 18	GCAGATATTGTTTGTGAACAGC	TACAATACTACACTGGGACCACC	405
Exon 19	CCAACCTGTGATATTCCTAGGAG	GTAACCCTCACGATACAGTCTTAGG	330
Exon 20	CAACTGAATTTGAAGAGGGCG	GGCAGGGGACACTTCAAGAG	233
Exon 21	GCATCCGAAAGAGACAGAGC	GGGATCTTGACATTTCTTACCTCC	333
Exon 22	CTTGTAGGTCCTTGTGTTCAACAG	GGAAAGTAAAGACTGTGGCAGG	414
Exon 23	CTTAGAGAAGCCAGCTTTTATGTG	GAAATACTCTATTGCAGTACCTCTGG	389

III. Primers used for the resequencing of targeted genes -cont'd.

BCL9	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 4	TGCAACCCGAGAGGAACT	TGTTCATGGTGTGCCTATGG	181
Exon 5	CCAATAGAAACTGCCTCTCTC	TGTCTGATATTAGGCTCTCTGG	496
Exon 6	AATTCTCAAGGTGGGTTTGTG	GATCAATCTGCAATGGGAATAAG	362
Exon 7	TCCCTTACAAGTTTATTGTGTG	CCATCAAGACTACCTTCCTATC	297
Exon 8_1	TATGTGCCTGGGGACAATTC	TCCAGCTGCTCCTGAGATAG	509
Exon 8_2	TACTGGGCCCAACTCAACTC	TCACTAGGGGTCATCTGGTATG	666
Exon 8_3	GATAGCGTGGCTGAAACTGC	TCCTGTTCATCTCCATGCTG	575
Exon 8_4	GGTCTTTCTGGAGTCAGTTGG	TTCTCTGAGATTCCGCAAGC	651
Exon 8_5	AGGAGATGCTGAAATTACGC	CAGGAAGTGATGGAGACTTG	549
Exon 8_6	GGAAGCCCTTGGATATATCTGTG	CCACAAATATGGCCATTTCAAC	509
Exon 9	CTCATTCCTGGCCTTGCTAG	CCTATTCAGGTGTATAGCCATTGA	451
Exon 10_1	CTTATTGATGGCAGGGATTG	CAGGACAGTGAAGGAGTCAG	607
Exon 10_2	AGGTTCCATTCCCTCACAATG	AGATTGTGCTGGTGACATCATC	569
Exon 10_3	CAGAATGGGACTAGCATTACC	TAGTAAGTACTCCTGGAACTGCC	524

ACP6	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 10	TGCTCTCTCCTCTTCCCAAAC	GGATGTAACTCGACTCTTTCTTCC	469
Exon 9	AGAGAGCTGACTGTGCCAAG	GTCTTACGTAAAGGAGGCAGG	373
Exon 8	CCTGGTTCTTGCCTCACTAAG	CTGTCTGGACTCATGAGGTCC	288
Exon 7	ATCAGCTGTCAGGTGTGAGTGG	AGATGGGCTCCTGCAGGTTC	295
Exon 6	GAGGCCTCCACCCACTTCTC	ATCCCAAATGATAACCCAAGTGTG	312
Exon 5	CCTCAGTATAGACAGAGTTCTGC	CTCTGCTCCATCTCTGTAGAC	266
Exon 4	TGCTGTGGATTATCACCAAG	CTCCTTATGTACTCTTGAAAGACC	587
Exon 3-2	TGAGAGACTCGAAGTGCAATG	GATTCCAGCCATGTGTTCTG	675
Exon 1	GCGAGTAAAGCTCTGAAGATGTG	CGCAGAAGACTTGTGTTTGC	375

GJA5	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 2_1	AGCATCAGTTCAGAAGGGAC	CTTAGCCTGGCTGAACTCTAC	499
Exon 2_2	CTCCAGGCACTGATTAAAGTCG	TGGCAGAGAAGGCAGAACTG	437
Exon 2_3	GATTCCGTAGATGAAGTACTGG	CAGACCTTCACAGAACATCC	621

GJA8	**Forward (5´-3´)**	**Reverse (5´-3´)**	**Size (bp)**
Exon 2_1	AGATATTGACTCAGGGTTGCATTG	AGCGGTACAGAGGCAGGATC	620
Exon 2_2	GGCAGCAAAGGCACTAAGAAG	CTCGCCCTCCACTTCTTGTG	603
Exon 2_3	GGGCTATCAGCTCCTAGAAGAAG	GTTGGCACCTTTTCCTTTCATC	581

III. Primers used for the resequencing of targeted genes -cont'd.

TCF4	Forward (5´-3´)	Reverse (5´-3´)	Size (bp)
Exon 19	AGCAATGTGGCAACTTGGAC	GGTCAGACACGCAAGAAGAG	437
Exon 18	GCTTGAAAGTCTACTGTCTGCC	TGTATTGTGAACTGTCAAAGCC	519
Exon 17	AGTTTCTTCCCGTTCTGTTC	AGTATGAATTGTCTGCTGGC	330
Exon 16	GAGGCTGGGTATCAACACTGG	CCAGCCTTCATCAGGTCCTC	294
Exon 15	GGCTCATCGTATGTTAAGTG	GGCAGTACCGTATATAGCAC	609
Exon 14	TCTTGGAGAGTAAAGGAGACTGAAC	ATCATTGCCATTGCCATTTC	256
Exon 13	GAGTTTCCACCTACAAAATCAGG	GAATTGCGTTGACAACTTCG	254
Exon 12	ATTGAGTGCAGCTTAGTACC	GAGTAAATGGACCAGGAATAG	370
Exon 11	TGGTTATGCTCAAGCCAGTG	TTGCCATTATATCCCTCCACAG	540
Exon 10	GCTGACATTGTGATAAAGAAGGC	CTTGGGAAGAAATGATGACTGTC	550
Exon 9	CATCGGCTCTAAGTGAAGCAGATC	AAGGGCACTGTTTCTAGGGTTTG	503
Exon 8	GCAATTAATCACTCCAATGTGG	CGTGGTAACTACAGTTTGCTAAGG	362
Exon 7	AGTTCTGCTACCATAAGCTG	TACTAACAATCAGTTGGGAGG	400
Exon 6	CAGCAGCGATCTAAAGATGAG	GAGCAGTAGATGTCTGTTACCTG	342
Exon 5	AACACAGACTGCCAATCCTC	TTGGTAGTGGAGCTATGTTTG	336
Exon 4	GTCAAGCTGATCCTCATTAAAC	AGCCACTATCATCATGGTAG	408
Exon 3	CAATAACCGTATGATTACAGGC	GTACTGCTTGGCCATCTAATG	443
Exon 2	CCTACTGGTTTCTAGCTGAAGTG	CCAGTCTCCAAAAATCCGATTG	254

CHD1L (cDNA)	Forward (5´-3´)	Reverse (5´-3´)	Size (bp)
Exons 4-10	GCGACAAGGAGGAAAGAGCCTGC	AAAGGCTCCGGCTCCACACCA	625

IV. Primers used in mass extension based genotyping. A total of 4 assays were used for genotyping of the variants selected for the follow-up of the *NRXN1*, *TCF4* and the 1q21.1 microdeletion loci. For each variant 3 primers were designed. 1st and 2nd primers were used for the amplification of the target region. UEP primers with their respective volume adjustments were used for the single base extension reaction in the iPLEX Gold Assay of Sequenom.

Assay	Variant ID	Gene	1st primer (5´-3´)	2nd primer (5´-3´)	UEP primer (5´-3´)	UEP conc. (µM)
1	rs45563244	*CHD1L*	ACGTTGGATGCATAATAGCAAGTGTTCATC	ACGTTGGATGCGAATAAGTCGCTCAGTACC	CCCACGTATTGGACATGC	7
1	18:52901933	*TCF4*	ACGTTGGATGGCCATCTTCACGATGGGTC	ACGTTGGATGTCTGGAAATAGCTGTCGCAC	TAAAGGGCGAGAGGAA	7
1	18:53018138	*TCF4*	ACGTTGGATGTACCTTACCCATGGCACTAC	ACGTTGGATGACCTGGTTCCCAGTACTATC	GGAGGCCTCCTTCGGG	7
1	18:52928751	*TCF4*	ACGTTGGATGGCCTGGTTTTCATATTCTGC	ACGTTGGATGTTTCCCCAGAGCATCTCCAG	TTTGCAGCAAATAGAGGAAG	9,13
1	18:52946804	*TCF4*	ACGTTGGATGAGCAGCAGCATCTTACCTTG	ACGTTGGATGCGACTACAATAGGGACTCGC	CATGAAGAAGGAGCTAGG	9,13
1	rs143555588	*TCF4*	ACGTTGGATGTCCATCTCTCTCAGGTAGTG	ACGTTGGATGCCACCTACAAAATCAGGAAG	TTCAAATCCCATTTCATCAT	9,13
1	rs144346949	*TCF4*	ACGTTGGATGCTTCTGATTAAAGTTCACCC	ACGTTGGATGAAAGCACTTGCTTCGGTGAG	TATATGGAGTCCAAAGTCA	9,13
1	rs189454938	*TCF4*	ACGTTGGATGATGGAAAGACAGAGGACGAG	ACGTTGGATGTGGTAAGGCAGACATACACG	GGACGAGGTTTAATCAAC	9,13
1	1:146740514	*CHD1L*	ACGTTGGATGACCACCTACCCAGAATACAG	ACGTTGGATGTTGGAGACCACCTGACTGAG	CCAGAATACAGGAATGCTAGTAG	11,6
1	1:146757200	*CHD1L*	ACGTTGGATGTGTCTGCGTTAACCAAGCTG	ACGTTGGATGGTGGTCTCACTATACAATAG	AAAACAGCTGGCGGCCACAGTT	11,6
1	rs140555192	*CHD1L*	ACGTTGGATGGGCCTAGAATGCAAAATAAG	ACGTTGGATGGAAGTGTGGGACAAAAGTAG	TTAAGTTCTCTAGGGTTCATTG	11,6
1	rs76956936	*TCF4*	ACGTTGGATGAAAGCTGCGTGTCTGAAAAG	ACGTTGGATGATGCGTCTCCCATTCCAGG	ATGCGTGTCTGAAAAGAAGGGA	11,6
1	1:146736063	*CHD1L*	ACGTTGGATGCCTTGCATTCCAGTGTTATG	ACGTTGGATGATGGGAGTTCCGGTCAACAG	TCCAGTGTTATGGAATTTTTATTTTA	14
1	1:146757032	*CHD1L*	ACGTTGGATGCCTGGTGATGGATCTTGTTC	ACGTTGGATGTCTTCTGGACTCAGAACCCG	GGGTCCAGGCCTTGTGGAGGGATCTA	14
1	rs142236750	*CHD1L*	ACGTTGGATGGAGAAGGAGACTCATAGAGG	ACGTTGGATGCTTGGTTAACGCAGACATAC	GGGCGACTCATAGAGGAGAAGAA	14
1	rs148434783	*CHD1L*	ACGTTGGATGCTGTCTGGCATTAAGATGGC	ACGTTGGATGGGTGGAAGAGCTTACCTTTC	GAAGAGAAGAGGGCCTGAAGAAG	14
1	18:52942992	*TCF4*	ACGTTGGATGGCCTCAGAAAGTCATGGTTG	ACGTTGGATGCATAGCCAGGCTGATTCATC	ATCATAAAGTCATGGTTGTCTTTGT	14
1	rs148658897	*TCF4*	ACGTTGGATGTTTTTCCCACTGCTCACAGG	ACGTTGGATGTCTTTGAGGAGCTCTGAAGG	GGGATCACAGGAGGTGAAAACATCT	14
1	rs191953257	*TCF4*	ACGTTGGATGTCCAGGTTTGCCATCTTCAG	ACGTTGGATGACTTCTATCTCCTTCCCATC	TTCTGCTTTGTATATTGCTTATTT	14
2	rs138570124	*TCF4*	ACGTTGGATGGCCACGCCATCTTCACGAT	ACGTTGGATGTCTGGAAATAGCTGTCGCAC	TCTTCACGATGGGTCC	7
2	rs35918540	*TCF4*	ACGTTGGATGGCAGAATATGAAAACCAGGC	ACGTTGGATGTTGTGAGTAAATGGACCAGG	AACCAGGCAGTGAGA	7
2	rs144512908	*CHD1L*	ACGTTGGATGCCTCACGATACAGTCTTAGG	ACGTTGGATGATGAGCTGGCTGGGAAAATG	CACCCTTAACCTTTCCCT	9,13
2	rs4950394	*CHD1L*	ACGTTGGATGTCCTAGAAGCAAGTCTGCTG	ACGTTGGATGTGGGCCAATTCTTAAGGCAC	AGTCTGCTGTCCTTCAT	9,13
2	rs143944746	*TCF4*	ACGTTGGATGCTACCTCTGTAAGGGTCCTG	ACGTTGGATGCCATTCTCTTCTGCCAAACC	GACAGGAAGCTGTGGAAC	9,13
2	rs147445499	*TCF4*	ACGTTGGATGTCACCGAAGCAAGTGCTTTC	ACGTTGGATGTGCAGCAAATAGAGGAAGCG	ACGTCTGGGAGCTGCCG	9,13
2	rs36008075	*CHD1L*	ACGTTGGATGTCTTCAGTCTCCTGTTGACC	ACGTTGGATGGAGATCAGGCTCCACAAAAC	CCCATCCAGAACAGCCTCCA	11,6
2	rs7547279	*CHD1L*	ACGTTGGATGTTGGATAAACTGCTGGCCTC	ACGTTGGATGACCCACTGGCCATCTTTTGT	CCAGAGACCTGGAGTCCAT	11,6
2	18:52928669	*TCF4*	ACGTTGGATGCTTCTGATTAAAGTTCACCC	ACGTTGGATGAAAGCACTTGCTTCGGTGAG	TCACCCTTTACAATGGTACAT	11,6
2	rs71083827	*CHD1L*	ACGTTGGATGCCCATGTGTATTTTATACCC	ACGTTGGATGCGTGGAAGATGAACACTTGC	TCAATTTTTGATGAGTGAAGAC	14
2	1:146736068	*CHD1L*	ACGTTGGATGATGGGAGTTCCGGTCAACAG	ACGTTGGATGCCTTGCATTCCAGTGTTATG	CCTGAGAACTCAAATACATAAATAAA	14
2	1:146737517	*CHD1L*	ACGTTGGATGAGAACCAAGGCAATGAACCC	ACGTTGGATGATCATACTGGCCTAGAATGC	ATGAACCCTAGAGAACTTAAA	14
2	rs185219867	*CHD1L*	ACGTTGGATGTGACCTCCAGGTATGATATG	ACGTTGGATGGAGTGCTGGTTCTTTTCATT	CTTCACTTTGGAATATACCAATA	14
2	rs2275250	*CHD1L*	ACGTTGGATGGCTTGCAAGTCATTCTGAGG	ACGTTGGATGTGTTTTCAGGTGGAGTTGGC	CAAGTCATTCTGAGGATTAAAGTC	14
3	rs148289715	*CHD1L*	ACGTTGGATGCCTTTCATTTTCCCAGCCAG	ACGTTGGATGTTTACAGCTCTGGAAAAGCG	CCCAGCCAGCTCATA	7
3	rs142335408	*FMO5*	ACGTTGGATGGACAGCTGCCATATTTGAGG	ACGTTGGATGGTCAAAGCTATAGCCTGTGG	ATTTGAGGATGGCTCC	7
3	rs58351438	*FMO5*	ACGTTGGATGCCCTCTGGGTTCTTATAGTC	ACGTTGGATGTATTTTCCCTTCTCTTGCAG	GTACTGCCCTTTGAACT	9,13
3	rs139791996	*CHD1L*	ACGTTGGATGAGAGCGAGCCAGAGGACCTT	ACGTTGGATGGAAGTAGCATCTGGGTCTTG	CAGAGGACCTTGAGAAT	9,13
3	1:146751866	*CHD1L*	ACGTTGGATGCCATTTGCTCTGCTCTAACC	ACGTTGGATGATGCCTTGCCTGCAGCAGAA	CTGCTCTAACCTCCAACTT	9,13
3	rs113139670	*CHD1L*	ACGTTGGATGACAGTGTCTCGACCAATCAG	ACGTTGGATGACCCTTATAGCACACTACCC	CCGAATAACTTTAACAGACC	11,6
3	rs144757186	*CHD1L*	ACGTTGGATGTGTAGTGGCCAAAAGGTGTC	ACGTTGGATGTCTCCCAAATGACCCAGATG	CACCTCTGTAATCCATATAGT	11,6
3	1:146658628	*FMO5*	ACGTTGGATGTGATGCGATCATCTGTGGTG	ACGTTGGATGATCCACTATCGTGTACAGGG	TCTGTGGTGAGGATAGCTTTTC	14
3	rs144288940	*CHD1L*	ACGTTGGATGCTTCTTGGGAAGCTCTGTAG	ACGTTGGATGCAAGTGAACTGCACAAACTC	TCTGTAGCTACCTCAGCTTTCACTC	14
4	rs112934082	*NRXN1*	ACGTTGGATGTTGGGCGGCTCATCGTCCAG	ACGTTGGATGTTCGTGACGTGAGGGTCAAC	GCCCGTGGACAGCGGCG	9,13
4	rs112638127	*NRXN1*	ACGTTGGATGTTGATCTTCTTGCAGGTGAC	ACGTTGGATGGGCATGTACAAGTTTTGGTA	TCTTTAGCTACTCCTCCTA	9,13

V. Association analysis of *CHD1L* gene exluding the single ungenotyped variant.

	CMC	WSS	RANK	DBM*	REP*	SKAT	C-alpha	MAF
CHD1L	0.279	0.242	0.232	0.255	0.132	0.528	1.000	0.01
	0.725	0.868	0.351	0.027	0.337	0.167	0.108	0.03
	0.511	0.746	0.346	0.042	0.421	0.198	0.297	0.05

The reported p-values are given for calculations based on 10000 permutations. Significant p-values below 5% significance level are given in bold. The better performing two- and one-sided test statistics for DBM and REP methods are given, respectively.

VI. Single marker analysis of *CHD1L* gene in the German sample. Empirical p-values are derived for the total [(a)] and the genotyping [(b)] samples by 1000 times permutated case-control labels for Pearson's chi-square test. The total sample includes the resequencing sample and the independent genotyping sample. The total allele counts for the total sample are 3976/4554 (patient alleles/control alleles). The total allele counts for the independent genotyping sample are 3792/4366 (patient alleles/control alleles). Variants are classified in the corresponding MAF categories based on their combined case/control frequencies in the combined sample. *The p-values are based on 10,000 permutations as the variants which yielded significant p-values below 10% significance level (p<0.1) in the initial analysis were re-analyzed with 10,000 times permutated case-control labels. MAF, minor allele frequency; syn, synonymous.

Genomic position (hg 19)	dbSNP ID	Effect	MAF	Minor allele counts (patients/controls)[a]	Empirical p-value	Minor allele counts (patients/controls)[b]	Empirical p-value
146736063	-	-	0.01	0/1	0.750	0/0	-
146736068	-	-	0.01	0/1	0.725	0/0	-
146736137	rs36008075	-	0.03	89/56	**5.00E-04***	85/55	**0.00095***
146737517	-	-	0.01	3/1	0.229	2/1	0.451
146737540	rs140555192	-	0.03	76/88	0.984	71/81	0.946
146740514	-	syn	0.01	4/5	0.865	3/4	0.868
146742648	rs144757186	D381N	0.01	1/0	0.252	0/0	-
146747069	rs2275250	syn	0.01	5/6	0.886	4/6	0.656
146747965	rs185219867	-	0.03	40/57	0.298	37/55	0.218
146751782	rs7547279	syn	0.05	120/156	0.384	118/149	0.542
146757032	-	T629N	0.01	0/2	0.367	0/1	0.732
146757132	rs142236750	syn	0.01	27/28	0.746	25/28	0.953
146757200	-	-	0.01	10/6	0.170	9/6	0.258
146758054	rs139791996	G700R	0.01	27/20	0.104	26/20	0.155
146759387	rs148289715	I765M	0.01	9/20	**0.080***	8/19	**0.067***
146759428	rs144512908	-	0.03	49/59	0.794	48/56	0.966
146765379	rs148434783	I827V	0.01	11/16	0.494	10/16	0.366
146766070	rs67589628	-	0.05	119/150	0.542	117/143	0.699
146766122	rs45563244	syn	0.03	113/134	0.880	108/131	0.908
146767149	rs4950394	S885A	0.05	118/150	0.506	116/143	0.675

VII. *In silico* analysis of splice site variants observed in *CHD1L* gene.

A) The splice site variation at chr1:146747766 effecting the acceptor site of exon 14

i) Wild type acceptor site

AGgtggagttggcatgaacttaacagcagcagatactgtgatttttgttgacagtgactttaatcctcagaatgacttgcaagcagctgcca
gggctcatcgcattggccaaaacaaGTA...............TACTTTTGCCCAGCTTTGCACACCCTTATAGCACACTACCC
TTGTTTGACTTATGTCCAGgtctgttaaagttattcggctgattggtcgagacactgtggaagaaatagtctataggaaagcagcc
tccaaactgcagctcaccaacatgatcatagaaggaggccattttactctgggagcccagaaacccgctgccgatgctgacctccagGT

ii) Upstream alternative acceptor site

AGgtggagttggcatgaacttaacagcagcagatactgtgatttttgttgacagtgactttaatcctcagaatgacttgcaagcagctgcca
gggctcatcgcattggccaaaac**aa**GTA...............TACTTTTGCCCAG**CTTTGCACACCCTTATAGCACACTACC**
CTTGTTTGACTTATGTCCGGgtctgttaaagttattcggctgattggtcgagacactgtggaagaaatagtctataggaaagca
gcctccaaactgcagctcaccaacatgatcatagaaggaggccattttactctgggagcccagaaacccgctgccgatgctgacctccagG
T

VII. *In silico* analysis of splice site variants observed in *CHD1L* gene-cont'd.

iii) Downstream alternative acceptor site

AGgtggagttggcatgaacttaacagcagcagatactgtgatttttgttgacagtgactttaatcctcagaatgacttgcaagcagctgcca gggctcatcgcattggccaaaacaaGTA................TACTTTTGCCCAGCTTTGCACACCCTTATAGCACACTACCC TTGTTTGACTTATGTCCGG**gtctgttaaagttattcggctgattggtcgagacactgtggaagaaatagtctataggaaag cagcctccaaactgcag**ctcaccaacatgatcatagaaggaggccattttactctgggagcccagaaacccgctgccgatgctgacctcc agGT

Exonic and intronic sequences are are given in lower and upper cases, respectively and they are always defined according to the wild type (i) splice site usage. The first and the second exons given in the figure are exon 13 and 14, respectively. The codons are denoted as alternating grey shaded and unshaded units. A phase 2 intron interrupts the reading frame between the second and third nucleotides of the codon depicted in red (i). This codon encodes the 462nd amino acid residue. i) The wild type donor site (GT) of exon 13 and the wild type acceptor (AG) site of exon 14 are shaded with yellow, ii) The wild type donor site (GT) of exon 13 and the alternative acceptor (AG) site of exon 14 located 45 bp upstream of the wild type acceptor site are shaded with yellow. The alternative acceptor site has a consensus value of 85.23 according to HSF. Alternative acceptor site usage leads to a frameshift event corresponding to amino acid position 462 (depicted in red letters) and leads to insertion of 16 amino acids encoded by the codons given in bold. The pre-mature stop codon corresponding to amino acid position 485 is shaded with green, iii) The wild type donor site (GT) of exon 13 and the alternative acceptor (AG) site of exon 14 located 77 bp downstream of the wild type acceptor site are shaded with yellow. The alternative acceptor site has a consensus value of 84.27 according to HSF. The sequences given in bold are skipped due to alternative splice site usage which leads to deletion of 26 amino acids. A frameshift event occurs at amino acid position 462 (depicted in red letters) and leads to a pre-mature stop codon shaded with green and corresponding to amino acid position 466.

B) The splice site variation at the at chr1: 146751866 effecting the donor site of exon 15

i) Wild type donor site

AGttgagtgagatactcaaatttggtttggataaactgctggcctctgaggggagcaccatggatgaaatagacctggagtccatactggg agaaacaaaagatggccagtgggtctctgatgccttgcctgcagcagaaggagggagcagagatcaagaggaaggaaGTAAGTTG GAGGT…………TAGaaaatcatatgtacttatttgaaggtaaagattattctaaagagcccagtaaggaagacagaaaatcatttgaa caactggtaaaccttcagaaaacccttttggagaaagctagtcaagagggccgatcactccgaaataaaggcagtGT

ii) Downstream alternative donor site

AGttgagtgagatactcaaatttggtttggataaactgctggcctctgaggggagcaccatggatgaaatagacctggagtccatactggg agaaacaaaagatggccagtgggtctctgatgccttgcctgcagcagaaggagggagcagagatcaagaggaagga**aGGAAGTTG GAG**GT…………TAGaaaatcatatgtacttatttgaaggtaaagattattctaaagagcccagtaaggaagacagaaaatcatttgaa caactggtaaaccttcagaaaacccttttggagaaagctagtcaagagggccgatcactccgaaataaaggcagtGT

Exonic and intronic sequences are are given in lower and upper cases, respectively and they are always defined according to the wild type splice site usage (i). The first and the second exons given in the figure are exon 15 and 16, respectively. The codons are denoted as alternating grey shaded and unshaded units. A phase 1 intron interrupts the reading frame between the first and second nucleotides of the codon depicted in red (i). This codon encodes the 569th amino acid residue. i) The wild type donor site (GT) of exon 15 and the wild type acceptor (AG) site of exon 16 are shaded with yellow, ii) The alternative donor site (GT) of exon 15 located 9 bp downstream of the wild type donor site and the wild type acceptor (AG) site of exon 16 are shaded with yellow. The alternative donor site has a consensus value of 90.88 according to HSF. Alternative donor site usage leads to a frameshift at amino acid position 569 (given in red) and to insertion of 4 amino acid encoded by the codons given in bold. The pre-mature stop codon corresponding to amino acid position 600 is shaded with green.

AVIII. Logistic regression analysis for the individual *CHD1L* variants.The genotyping is performed in a large case-control cohort composed of individuals of German, Dutch and Danish origin. The logistic regression analysis was performed to control for different ethnicities. The empirical p-values were calculated both for the total sample [(a)] where the initial and extended discovery samples were included and the independent genotyping sample [(b)]. Empirical p-values were derived from 1000 permutations of case-control status where the specific case/control ratios of each population were kept constant. Minor and total allele counts of variants are given separately for each population based on the respective post QC- genotype calls of each individual variant; (n_{pat}/n_{cont})G= Germany, (n_{pat}/n_{cont})D= Denmark, (n_{pat}/n_{cont})N= the Netherlands.

Genomic position (hg19)	dbSNP ID	Effect	Minor allele counts[a] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Total allele Counts[a] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Empirical p-value	Minor allele counts[b] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Total allele Counts[b] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Empirical p-value
146737632	rs144288940	R261*	3/0 0/1 1/0	3992/4560 3734/3692 1292/820	0.195	2/0 0/1 1/0	3616/4372 3734/3692 1292/820	0.402
146742648	rs144757186	D381N	1/0 0/0 0/0	3992/4560 3734/3692 1292/818	0.374	0/0 0/0 0/0	3616/4372 3734/3692 1292/818	-
146747766	rs113139670	splice site	7/7 8/8 5/4	3992/4560 3734/3688 1292/820	0.946	6/7 8/8 5/4	3616/4372 3734/3688 1292/820	0.933
146751866	-	splice site	1/1 0/0 0/0	3992/4560 3742/3692 1292/822	0.372	0/1 0/0 0/0	3616/4372 3742/3692 1292/822	0.756
146758054	rs139791996	G700R	27/20 14/20 4/3	3990/4560 3732/3688 1292/812	0.715	24/20 14/20 4/3	3614/4372 3732/3688 1292/812	0.871
146759387	rs148289715	I765M	10/20 6/5 3/2	3992/4560 3740/3688 1292/822	0.262	9/19 6/5 3/2	3616/4372 3740/3688 1292/822	0.308

AIX. Logistic regression analysis for the individual *FMO5* variants. The genotyping is performed in a large case-control cohort composed of individuals of German, Dutch and Danish origin. The logistic regression analysis was performed to control for different ethnicities. The empirical p-values were calculated both for the total sample (a) where the initial and extended discovery samples were included and the independent genotyping sample (b). Empirical p-values were derived from permutations of case-control status where the specific case/control ratios of each population were kept constant. Significant p-values below 5% significance level are given in bold. The reported p-values for R319G are given for calculations based on 10000 permutations and for the other variants for calculations based on 1000 permutations as the analyses were re-ran with 10000 permutations only if significant p-values were derived from the initial analysis. Minor and total allele counts of variants are given separately for each population based on the respective post QC- genotype calls of each individual variant; (n_{pat}/n_{cont})G= Germany, (n_{pat}/n_{cont})D= Denmark, (n_{pat}/n_{cont})N= the Netherlands.

Genomic position (hg19)	dbSNP ID	Effect	Minor allele counts[a] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Total allele Counts[a] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Empirical p-value	Minor allele counts[b] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Total allele Counts[b] (n_{pat}/n_{cont})G (n_{pat}/n_{cont})D (n_{pat}/n_{cont})N	Empirical p-value
146684095	rs58351438	K166E	23/33 28/36 12/7	3992/4560 3736/3690 1292/822	0.262	18/32 28/36 12/7	3616/4372 3736/3690 1292/822	0.149
146672962	rs142335408	R319G	11/5 4/1 2/1	3990/4560 3740/3690 1292/818	0.024	10/5 4/1 2/1	3614/4372 3740/3690 1292/818	0.053
146658628	-	R485*	1/0 0/0 0/0	3992/4560 3740/3692 1292/822	0.438	0/0 0/0 0/0	3616/4368 3740/3692 1292/822	

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to my supervisor Prof. Dr. Sven Cichon for his continuous scientific and moral support as well as his guidance throughout my PhD study which enabled this thesis.

I would also like to thank Prof. Dr. Michael Hoch for taking over the co-supervision of this work.

I deeply appreciate Prof. Dr. Markus Nöthen's contribution to this work through the valuable scientific discussions and the continuous support he provided. I am also grateful to him for reviewing parts of this thesis.

This study would not have been possible without the contributions of national and international collaboration partners. I would like to thank Prof. Dr. Marcella Rietschel, Prof. Dr. Dan Rujescu, Prof. Dr. Rainald Mössner, Prof. Dr.Wolfgang Maier, Prof Dr. Susanne Moebus from Germany, Prof. Dr. Roel A. Ophoff and the Genetic Risk and Outcome of Psychosis (GROUP) Investigators from Netherlands, Prof. Dr. Preben B. Mortensen, Prof Dr. Anders D. Børglum, Prof. Dr. David M. Hougaard from Denmark and their respective teams for the recruitment of the patient and control cohorts and for making these samples available to us for the conductance of this work. I would like to thank Prof. Dr. Christoph Lange for the collaboration on the statistical analyses. I express my deep gratitude to Dr. Heide Fier for the great fun times we had together analyzing and interpreting the genotyping data and for making statistics interesting to me.

I acknowledge the European Union for the Marie Curie fellowship I was granted under the EUTwinsS network (RTN, FP6) which financially supported me and this work.

I express my appreciation for the support and motivation my current supervisor Prof Dr. Regina Betz has given for the finalization of this thesis.

I would like to thank Bärbel Lippke, Nadine Fricker, Peter Tessman and Sandra Barth for providing the greatest technical assistance whenever needed. I am grateful to Dr. Andrea Hofman for her help with the analysis of the microarray data. Stefan Herms, this work would not have been possible without your contributions to data handling; I owe you a big 'Thank You'! It was much fun to work together with Andreas Forstner to whom I could always rely on. I would like to thank him for reviewing parts of this thesis and for being the great lab partner he is. I would like to thank all my current and former colleagues from Institute of Human Genetics for the fruitful scientific discussions, great working environment and above all for having made me feel home all the years I have been in Germany.

I am most deeply grateful to my colleague and dear friend Dr. Stefanie Heilmann who also reviewed parts of this thesis and gave much appreciated feedback and was always there for me with her valuable emotional support.

I would like to express my gratitude to my dear colleague and friend Gunhild van Loon for always being there for me in every possible way, starting from the first day I stepped into the Institute of Human Genetics.

I would like to thank Christoph Leitsch for the support he has given me during my studies. I am deeply grateful to Demet Toraman, Sinem Hacioglu, Melis Akman and Nilufer Erkal for always being there for me at both difficult and joyful times during the course of this thesis. I would like to thank my entire family; aunts, uncles and cousins, for their continuous and very valuable emotional support during my studies. I cordially thank Murat Uenalan for the endless motivation and support he has been giving me at all times.

There are no words which would appropriately express my gratitude to my parents Secil and Erman Basmanav. You have been there for me throughout my life always with endless care, support and love. You have always given me strength and relief at most difficult times. Thank you for being such great parents. I dedicate this PhD thesis to you.