A genetic, epigenetic and transcriptomic study of 22q11.2 Deletion Syndrome and its schizophrenia phenotype

Thesis submitted for the degree of Doctor of Philosophy

at the School of Medicine, Cardiff University (2017)

By

Thomas Monfeuga

Supervised by: **Professor Nigel Williams**
MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University

and: **Professor Meng Li**
Neuroscience and Mental Health Research Institute, School of Medicine, Cardiff University

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …… (candidate) Date…….

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD

Signed …… (candidate) Date…….

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …… (candidate) Date…….

Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date…….

Statement 4: previously approved bar on access

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans **after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.**

Signed …… (candidate) Date…….

Statement of contribution

Part of the data, work and results presented in this thesis have been generated/performed/analysed by different individuals. These instances are detailed in the relevant sections of the thesis and are recapitulated here.

Chapter 2: Schizophrenia polygenic risk score analysis in 22q11.2 deletion syndrome patients (p.31).

22q11.2 IBBC cohort samples have been provided by the International 22q11.2 Brain Behavior Consortium (raw .CEL files). The imputation of the 22q11.2 IBBC cohort data has been performed by Dr. Leon Hubbard (Cardiff University) (Genotype calling, pre- and post-imputation quality control and processing of the data were performed by the writer of this thesis). The CLOZUK data has been provided after imputation by Dr. James Walters (Cardiff University).

Chapter 3: Effect of the 22q11.2 deletion on DNA methylation (p.49).

Samples from the two cohorts' groups (Cardiff University's ECHO cohort, led by Prof. Marianne van den Bree and Prof. Michael Owen; and King's College London's BBAG study, led by Dr. Michael Craig) were provided after DNA extraction by the MRC Core team of the Centre for Neuropsychiatric Genetics and Genomics (Cardiff University). 450K DNA methylation arrays were prepared and run by Dr Afnan Salaka and Catherine Bresner (Cardiff University); EPIC DNA methylation arrays were prepared and run with the help of Catherine Bresner.

Chapter 4: CRISPR/Cas9 genome editing of DGCR8 in human embryonic stem cells (p.81).

DNA sequencing was performed by GATC Biotech.

Chapter 5: Transcriptome analysis of DGCR8 knock-out lines (p.106).

The induced pluripotent cell lines have been generated by Dr. Craig Joyce (Cardiff University) from dermal fibroblasts originated from skin biopsies provided by the National Centre for Mental Health (NCMH). Cell culture and RNA extraction in "Experiment 2" was done collaboratively with Matthieu Trigano (Cardiff University) (but different cell lines and experiments were analysed by Matthieu Trigano and the writer of this thesis independently). Post-extraction RNA integrity was assessed by Dr. Amanda Redfern (Central Biotechnology Services, Cardiff University). Sample preparation and sequencing was performed by Joanne Morgan (NGS Co-ordinator, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University). Alignment of RNA-sequencing reads was performed by Daniel Cabezas De La Fuente (Cardiff University).

Acknowledgements

First of all, I would like to thank my primary supervisor, Prof. Nigel Williams, for his guidance and his support throughout this PhD. I really appreciated the numerous discussions we have had during these 3 years; they were always constructive and interesting and will, I am sure, have a great impact for the rest of my scientific career. I would also like to thank my second supervisor, Prof. Meng Li, for all her help and feedback on my work.

I am very grateful to the Wellcome Trust for funding my project and offering me this great opportunity.

I would like to thank everyone that helped me during my PhD, who are probably too many to name here. I would like in particular to express my gratitude to Dr. Leon Hubbard, for all his assistance in bioinformatics. I am also thankful to everyone in the Li Lab for teaching me all I know about stem cell culture and being wonderful colleagues and friends; in particular it was a pleasure to work closely with my friend Matthieu Trigano during this PhD. I would also like to thank the people that helped me in various way in the last three years, especially Catherine Bresner and Dr. Afnan Salaka for generating the DNA methylation array data and helping me with the project, as well as Dr. Antonio Pardinas, Mark Einon, Prof. Peter Holmans and Dr. Elliot Rees for their help in bioinformatics and statistics. I am also grateful to all the researchers who collected the biological samples and generated the data analysed in this thesis and granted me permission to use it; and I am grateful to the patients who provided these samples. Last but not least, I want to thank everyone in the core team of the MRC Centre for Neuropsychiatric Genetics and Genomics for everything they taught me and their invaluable assistance.

Then I would like to thank my friends, who are an essential part of who I am. In Cardiff, of course, Matthieu, Michal, Liad, Jamie and everyone else that I have met since I arrived in the UK; but also every friend I have made in Grenoble, Lyon, Marseille, Boston and all around the world. Most of all, I want to thank Peggy, for always being there for me.

Finally, I want to thank my family. They got me to where I am now; and despite the distance, their support was always crucial for me. I dedicate this thesis to my grand-parents. I know they are or would have been proud of me and I cherish every moment I had with them.

Abstract

22q11.2 Deletion Syndrome (22q11.2DS) is a genetic disorder that results from a hemizygous deletion at chromosome 22q11.2, occurring at an incidence of 1 in 4000 live births. It is associated with a wide range of clinical features, such as congenital heart disorders and abnormal facial features. 22q11.2DS patients also have an increased risk of neuropsychiatric disorders, with deletions at 22q11.2 being the highest known risk factor for schizophrenia. However, the mechanisms underlying 22q11.2DS symptomatic variability are still unclear. This thesis addresses this issue by investigating genetic, epigenetic and transcriptomic changes related to 22q11.2DS.

Firstly, by using a polygenic risk score profiling approach it shows that the increased risk of schizophrenia in 22q11.2DS patients is partly due to an increased burden of common genetic variants associated with this neuropsychiatric disorder.

This thesis also presents evidence that DNA methylation, an epigenetic mark, is altered in 22q11.2DS patients compared to a control population that do not carry a deletion at 22q11.2. Microarray-based whole epigenome analysis showed that these patients have an altered DNA methylation profile that affects genes and biological pathway relevant to schizophrenia.

Finally, the CRISPR/Cas9 genome editing technology has been employed in human embryonic stem cells to delete one of the genes spanned by the 22q11.2 deletion. This gene, DGCR8, is a major component of the microRNA biogenesis pathway that is involved in the regulation of gene expression. The knock-out cell lines generated in this study were differentiated into neural progenitor cells to investigate transcriptome changes due the deletion of this gene during neurodevelopment.

In conclusion, this thesis shows that the increased risk for schizophrenia in 22q11.2DS patients depends in parts on common genetic variants located outside of the deletion. Moreover, different mechanisms involved in genetic regulation (DNA methylation, microRNAs) can possibly modulate the schizophrenia phenotype by affecting relevant genes and pathways.

Contents

Abbreviations

Abbreviation	Description
µM	Micromolar
22q	22Q11.2DS
22q11 IBBC	International 22Q11.2 Brain Behaviour Consortium
22q11.2DS	22Q11.2 Deletion Syndrome
450K array	Illumina Infiniumhumanmethylation450 Beadchip Array
5mC	5-Methylcytosine
Ab	Antibody
aCGH	Array Comparative Genomic Hybridization
ADHD	Attention-Deficit/Hyperactivity Disorders
ASD	Autism Spectrum Disorders
BBAG	Brain, Behaviour And Genetics in 22Q11.2 Deletion Syndrome Cohort
Bcell	B-Cell
Beta	Beta-Value
bp	Base Pair
BP	Biological Process
BPD	Bipolar Disorder
C	Cytosine
CAM	Cell Adhesion Molecule
CC	Cellular Component
CD4T	CD4+ T-Cell
CD8T	CD8+ T-Cell
CGI	CpG Island
CHD	Congenital Heart Disorder
CI	Confidence Interval
CLOZUK	CLOZUK Cohort
CNV	Copy Number Variant
COGS	COGS Cohort
CpG	Cytosine followed by Guanine
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DDR	DNA-Damage Response
DEPC	Diethyl Pyrocarbonate
DGCR	DiGeorge Syndrome Chromosomal Region
DMP	Differentially Methylated Probe
DMR	Differentially Methylated Region
DNA	Deoxyribonucleic Acid
DNAm	DNA Methylation
DNMT	DNA Methyltransferase
DSB	Double Strand Break
dsRBD	Double-Strand RNA Binding Domain
E13.5	Embryonic Day 13.5

ECHO	ExperienCes of people witH cOpy number variants Cohort
ECM	Extracellular Matrix
EPIC array	Infinium Humanmethylationepic Beadchip Array
eQTL	Expression Quantitative Trait Locus
ESC	Embryonic Stem Cell
EWAS	Epigenome-Wide Association Study
FACS	Fluorescence-Activated Cell Sorting
FC	Fold Change
FDR	False-Discovery Rate Corrected P-Value
FISH	Fluorescence In Situ Hybridisation
G	Guanine
gDNA	Genomic DNA
GEO	Gene Expression Omnibus
GFR	Growth Factor Reduced
GO	Gene Ontology
Gran	Granulocyte
gRNA	Guide RNA
GSE	GEO Series Record
GWAS	Genome-Wide Association Study
H7	WA07 Human Embryonic Stem Cell
HWE	Hardy-Weinberg Equilibrium
HDR	Homology-Directed Repair
hESC	Human Embryonic Stem Cell
Het	Heterozygous
hiPSC	Human Induced Pluripotent Stem Cell
Hom	Homozygous
IAA	Interrupted Aortic Arch
IBBC	International 22Q11.2 Brain Behaviour Consortium
IBD	Identity-By-Descent
INDEL	Insertion/Deletion Mutation
IPA	Ingenuity Pathway Analysis
iPSC	Induced Pluripotent Stem Cell
IQ	Intelligence Quotient
KO	Knock-Out
L1	Long Interspersed Nucleotide Element-1
LCL	Lymphoblastoid Cell Lines
LCR	Low-Copy Repeats
lncRNA	Long-Non-Coding RNA
log	Natural Logarithm
log10	Logarithm Base 10
LRES	Long-Range Epigenetic Silencing
LTP	Long-Term Potentiation
Mb / Mbp	Mega Base Pair
MDD	Major Depressive Disorder
MeDIP-seq	Methylated DNA Immunoprecipitation Sequencing

mESC	Mouse Embryonic Stem Cell
Met	Methionine
MF	Molecular Function
MHC	Major Histocompatibility Complex
microRNA-seq	MicroRNA Sequencing
min	Minute
miR	MicroRNA
miRNA	MicroRNA
MLPA	Multiplex Ligation-Dependent Probe Amplification
Mono	Monocyte
mQTL	Methylation Quantitative Trait Loci
MRC	Medical Research Council
mRNA	Messenger RNA
mRNA-seq	Messenger RNA Sequencing
mtRNA	Mitochondrial RNA
NAHR	Non-Allelic Homologous Recombination
NCMH	National Centre For Mental Health
NHEJ	Non-Homologous End-Joining
NK	Natural Killer Cell
NMJ	Neuromuscular Junctions
NonAffect	22Q11.2 Deletion Syndrome Individuals Without Psychotic Disorders
NPC	Neural Progenitor Cell
NLS	Nuclear Localisation Signal
OCD	Obsessive-Compulsive Disorder
OR	Odd Ratio
p	P-Value
PBMC	Peripheral Blood Mononuclear Cell
PBS	Phosphate-Buffered Saline
PBST	PBS + 0.3% Triton X-100
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Parkinson’S Disease
PFC	Prefrontal Cortex
PGC	Psychiatric Genomics Consortium
P_{perm}	Permutation P-Value
pre-miRNA	Precursor MicroRNA
pri-miRNA	Primary MicroRNA
PRS	Polygenic Risk Score
pT	P-Value Threshold
QC	Quality Control
qPCR	Real-Time Quantitative Polymerase Chain Reaction
Q-Q plot	Quantitative-Quantitative Plot
R^2	Correlation Coefficient
R^2 (PRS analysis)	Nagelkerke's Pseudo-R2
Rhed	RNA-Binding Heme Domain

RIN	RNA Integrity Number
RNA	Ribonucleic Acid
RISC	RNA-Induced Silencing Complex
RNA-seq	RNA Sequencing
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	Ribosomal RNA
s.d.	Standard Deviation
SCZ	Schizophrenia
SD	Segmental Duplications
snoRNA	Small Nucleolar RNA
SNP	Single-Nucleotide Polymorphism
SNV	Single-Nucleotides Variant
Stouffer(FDR)	Stouffer Combined FDR Value
TA	Truncus Arteriosus
TALEN	Transcription Activator-Like Effector Nuclease
TET	Ten-Eleven Translocation Family Of Proteins
TF	Tetralogy Of Fallot
TSS	Transcription Start Site
UTR	Untranslated Region
Val	Valine
VCFS	Velo-Cardio-Facial Syndrome
VSD	Ventricular Septal Defects
WS	Williams Syndrome
WT	Wild-Type

1. General Introduction

1.1. Overview and nomenclature

22q11.2 deletion syndrome (22q11.2DS) is a genetic disorder caused by a hemizygous deletion of chromosome 22q11.2. In the past it has been also designated by other names, such as DiGeorge syndrome, Velo-Cardio-Facial syndrome (VCFS), CATCH 22 syndrome or conotruncal anomaly face syndrome [1]. The different syndromes associated with 22q11.2 deletion and their date of discovery are described Table 1-1. 22q11.2DS syndrome patients can present with a wide range of diverse symptoms, for instance congenital heart disorders, facial dysmorphism, immune disorders as well as cognitive and neuropsychiatric disorders. Robin and Schprintzen reported in 2005 more than 180 clinical features associated with this syndrome [2]. The presence of these diverse symptoms is variable and lead to the use of different syndrome names depending on the diagnosis established from these features. The advances in genetic testing allowed the unification of these syndromes and facilitated the identification of patients who have the symptoms associated with the syndrome but do not carry the 22q11.2 deletion. For instance, a study by Driscoll *et al.* in 1993 identified that the deletion was detected in only 83% of patients diagnosed with DiGeorge syndrome and 68% of those diagnosed with VCFS in a cohort of 76 patients [3]. Consequently, 22q11.2 deletion syndrome is now only designated to patients that have this copy number variant (CNV) mutation, regardless of exhibited symptoms. For this reason, throughout this thesis only the denomination 22q11.2 deletion syndrome (22q11.2DS) is used.

Event	Date
Congenital thymic hypoplasia associated with hypocalcemia	1959
DiGeorge syndrome described	1972
Takao syndrome described (conotruncal anomaly face syndrome)	1976
Velocardiofacial syndrome described (Shprintzen syndrome)	1978
DiGeorge syndrome speculatively linked to chromosome 22	1981
Partial monosomy of chromosome 22 described	1982
"CATCH-22 syndrome" described	1989
Cayler syndrome associated with del22q11.2	1994

Table 1-1. Historical overview of the syndromes associated with the 22q11.2 deletion. Table adapted by permission from Springer Nature: Genetics in Medicine, De Decker 2001 [489], copyright (2001).

1.2. Epidemiology

Due to the variability of symptoms observed in 22q11.2DS patients and the absence of systematic genetic testing, the prevalence rates reported in the past were variable and affected by ascertainment bias. Birth defects are the most common symptom leading to a diagnosis of 22q11.2DS. Children born in the United Kingdom in 1993 with congenital disorders (mainly significant heart diseases) have been screened for the presence of the deletion by Goodship *et al.* in 1998 [4]. It established a prevalence of a deletion at 22q11 of 25.7 per 100,000 live births (around 1:3891). An ascertainment bias has been demonstrated in a Swedish population study [5]. They determined an incidence of 18.1:100,000 (1:5524) live births for the Western Götaland region and an incidence of 23.4:100,000 (1:4273) in the city of Gothenburg where a team of specialists of 22q11.2DS was posted. A 2016 systematic review of birth prevalence in Belgium, France, Singapore, Sweden, the United Kingdom and the United States of America reported 156 cases for 1,111,336 live births, with per-study rates ranging from 1:4525 to 1:9805 [6]. When restricting to cohorts of patients with heart diseases, the range was between 1:4000 and 1:7092 live births. No significant difference was observed when comparing prevalence rates between different racial groups within populations [6]. The commonly discussed prevalence rate in 22q11.2DS is 1 in 4000 live births. The incidence of 22q11.2DS cases in miscarriages has been estimated to be 1:1497, which is about three time higher than reported rates in the general population [7].

The 22q11.2 deletion is typically a *de novo* mutation, but parental transmission has been estimated to occur in between around 6 to 17% [8,9]. There is a significant difference in the parental origin of the deletion, with around 56-58% of maternal transmission [10,11].

1.3. The 22q11.2 deletions

Multiple hemizygous deletions have been characterised at the cytoband 22q11.2. The most frequent deletions in 22q11.2DS is around 3 Mb long and is shared by around 85 to 90% of the patients. The second most frequent deletion (5-10% of patients) is a smaller deletion of around 1.5 Mb which is nested within the 3 Mb deletion. Thus, the majority of 22q11.2DS patients share a common 1.5 Mb deletion. Some patients (2-4%) present more atypical deletions, such as central deletions. [12–17] (Figure 1-1.B).

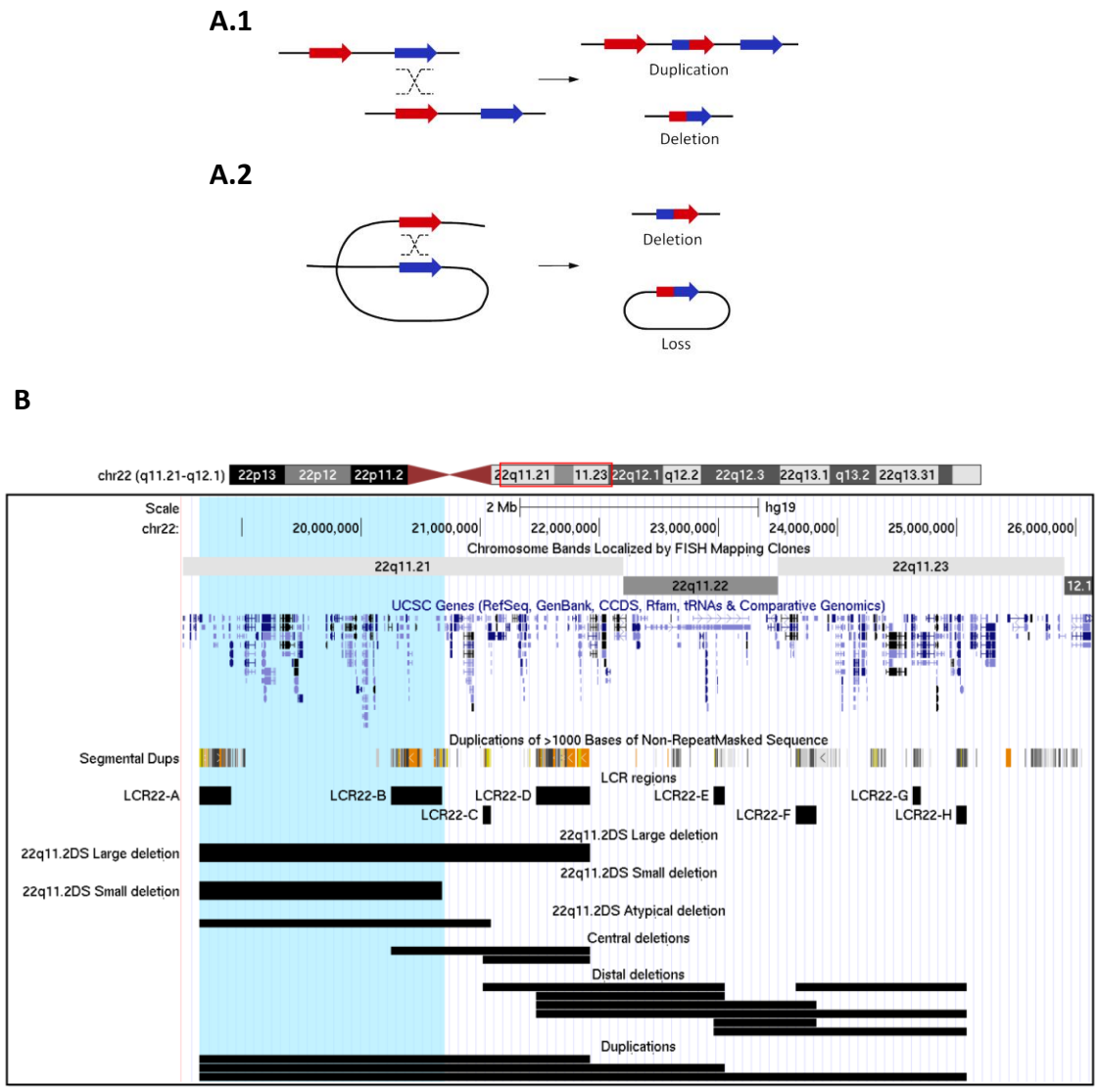


Figure 1-1. Genomic rearrangements in 22q11.2DS. (**A**) N**on-allelic homologous recombination (NAHR) events between paired low-copy repeats (LCRs)/segmental duplications (SDs).** Paired LCRs/SDs are depicted as bold arrows (red and blue) with the orientation indicated by arrowheads. (A.1) The inter-chromatid NAHR events between directly oriented LCRs/SDs result in deletions and duplications. (A.2) The intra-chromatid NAHR events between directly oriented LCRs/SDs can generate deletions and ring-shaped DNA segments that will be lost in subsequent cell divisions. Figure and legend adapted from Chen 2015 [490], used under CC BY-NY 3.0 license. (**B**) **22q11.2 region breakpoints and recombination events**. The 22q11.2DS LCR A to LCR B and LCR A to LCR D deletions are also referred as proximal deletions. Not all types of atypical deletions are represented. The region delimited by the LCR A to LCR B deletion is highlighted in light blue. Source: UCSC Genome Browser [491]. Deletions and duplication events as described in Torres-Juan 2007 [24] and Burnside 2015 [17]. LCR regions coordinates are as defined in Mikhail 2014 [22] (GRCh37/hg19).

The deletions have been shown to be delimited by breakpoints occurring in regions that contains low-copy repetitive (LCR) sequences. These regions are also flanked by inverted repeats [18]. The 22q11.2 region contains eight LCR regions (LCR A-H, as shown

Figure 1-1.B). Low-copy repeats are regions of the genome whose sequence share high degree of similarity and act as hotspots for non-allelic homologous recombination (NAHR) (as reviewed in Gu 2008 [19]). During meiosis and mitosis, alignment of these highly similar sequences can occur and cause recombination events that will lead to deletion or duplication of the regions surrounded by LCRs (Figure 1-1.A). The 22q11.2 region is affected by many NAHR events due to the presence of 8 LCRs. The common ~1.5Mb deletion occurs between LCR A and LCR B, while the larger ~3Mb deletion occurs between LCR A and LCR D (Figure 1-1.B). A small percentage of 22q11.2DS patients (around 2%) have a deletion of around 2Mb that spans the region from LCR A to LCR C [20]. These 3 deleted regions are also called proximal deletions. Central deletions have also been reported to span LCR B/C and LCR D [17]. There are also deletions that occurs outside of these regions and that are not part of 22q11.2DS. 22q11.2 distal deletion syndrome is characterised by the presence of deletions of different lengths nested within LCR C and LCR G [21,22]. Finally, 22q11.2 duplication syndrome is defined by a duplication of regions including LCR A to LCR D and that stop at either LCR D, LCR E or LCR G [23,24].

Interestingly, it has been observed in a small sample study that familial (inherited) 22q11.2 deletions are significantly enriched for the smaller A-B deletion rather than the usually more frequent A-D deletion [25].

1.3.1. 22q11.2DS molecular diagnosis

Early detection of 22q11.2 deletions implicated in 22q11.2DS was performed by fluorescence in situ hybridisation (FISH) [4,26]. Cytogenetic testing provides accurate results but is time consuming, and the extent of the deletion needs to be investigated by the use of multiple probes. Moreover, the probes might miss more atypical deletions. This method is still used but does not allow more high-throughput approaches. Multiplex ligation-dependent probe amplification (MLPA) has also been successfully used to detect the presence and the size of the deletion [27,28]. Using this method, a high-definition MLPA 22q11 kit has been developed to detect CNVs at 37 loci included within the region deleted in 22q11.2DS and more distal regions [28]. MLPA has been used to detect deletions with perfect sensitivity and specificity in dried blood spot samples from neonatal screening, which shows its potential for early diagnosis [29]. Other fast and cost-effective methods that have been used include real-time quantitative polymerase chain reaction (qPCR) and

multiplexed qPCR [30,31]. More advanced methods that have been employed for accurate detection of 22q11.2 deletions are array comparative genomic hybridization (aCGH) [32] and next-generation sequencing. The latter has been used to detect the deletion from maternal plasma [33]. Furthermore, these methods allow to screen the entire genome and assess the presence of other genomic aberrations than the deletion at 22q11.2.

1.4. 22q11.2DS symptoms

22q11.2DS patients can present a large range of different symptoms affecting many organs (Figure 1-2). The characteristic features of 22q11.2DS patients used to be referred as CATCH-22, as defined by Wilson *et al.* (1993): "Cardiac defects, Abnormal facies, Thymic hypoplasia, Cleft palate, and Hypocalcaemia resulting from 22q11 deletions" [34]. However, it does not describe well the important heterogeneity of clinical features present in 22q11.2DS patients, and ignore the high incidence of neuropsychiatric symptoms observed. The main symptoms observed in 22q11.2DS below are reported below; a more extensive list of clinical feature features can be found in Robin and Schprintzen (2005) [2].

1.4.1. Non-neuropsychiatric/neurological symptoms

1.4.1.1. Congenital heart disorders

One of the most common and well-characterised symptom in 22q11.2DS is the presence of congenital heart disorders. Around 80% of children born with 22q11.2 deletion have cardiovascular anomalies (ranging from 64 to 87% depending on studies) [35,36]. The most frequent (23%, range: 13-39%) of these anomalies is the Tetralogy of Fallot (TF), followed by Ventricular Septal Defects (VSD) (14%, range: 13-18%), Interrupted Aortic Arch (IAA) (10%, range: 4-16%) and Truncus Arteriosus (TA) (6%, range: 0.5-10%) [35]. Prenatal diagnose of congenital heart disorders in foetuses positively screened for 22q11.2 deletion found that only 4% of individuals presented a normal cardiac anatomy [37]. However, 65% of all pregnancies of 22q11.2DS foetuses resulted in a termination of pregnancy, which would explain the difference in observed rates of cardiovascular anomalies.

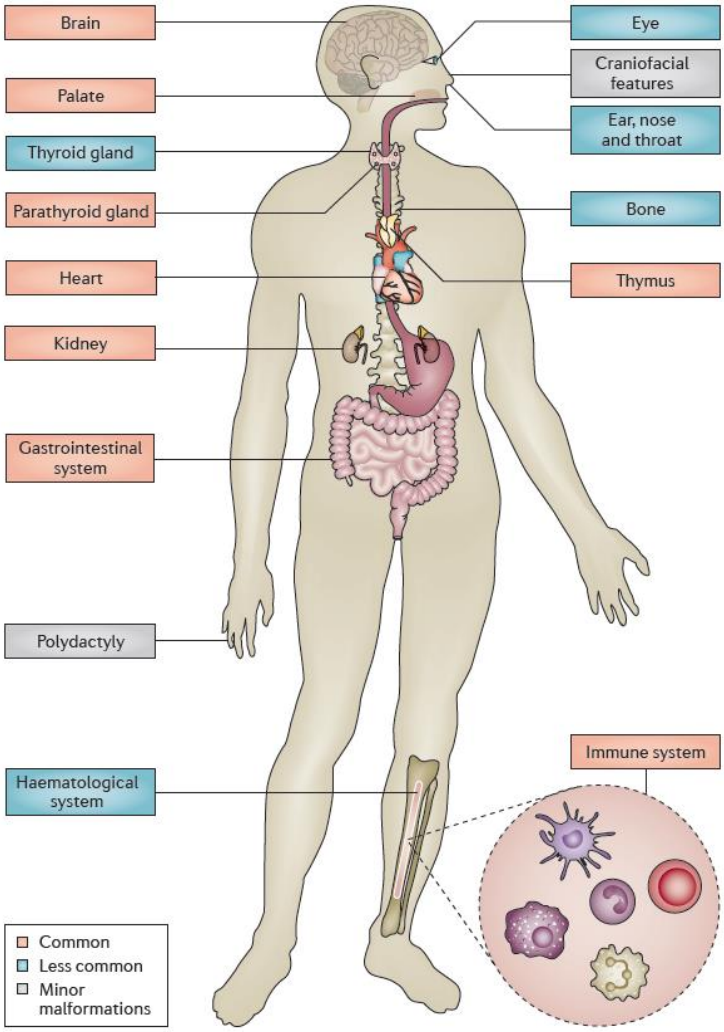


Figure 1-2. Organ and system involvement in 22q11.2 deletion syndrome. Figure reprinted by permission from Springer Nature: Nature Reviews Disease Primers, McDonald-McGinn 2015 [492], copyright (2015).

1.4.1.2. Craniofacial features

Facial dysmorphism is present in nearly all children with 22q11.2DS. The most commonly observed features are malar flatness, hooded eyelids, broad nasal bridge and/or tip, round/broad ears with thick/overfolded helix [38]. Attempts at using these facial characteristics as a diagnosis tool by specialists resulted however in relatively low specificity of 22q11.2DS diagnosis (51%, and 70% sensitivity) [39]. An important proportion of 22q11.2DS patients also present palatal anomalies. The most frequent are velopharyngeal insufficiencies (27-32%) and/or a cleft palate (9-11%) [40,41]. Finally, 14% of 22q11.2DS patients have an asymmetric crying facies (due to the a unilateral hypoplasia of one facial muscle, the depressor anguli oris) [42].

1.4.1.3. Immune and endocrine disorders

The abnormal development (hypoplasia) or complete absence (aplasia) of the thymus is relatively common in 22q11.2DS. Thymus hypoplasia is seen in around 80% of 22q11.2DS and leads to T-cell deficits (lymphopenia) [43,44]. Thymus aplasia and its resulting severe lymphopenia are however rare (0.5%) [45]. Recurrent infections as well as autoimmune diseases (*e.g.* juvenile rheumatoid arthritis or idiopathic thrombocytopenia purpura) have been observed in 22q11.2DS patients but do not seem to be correlated with the lymphocyte counts [43].

Around 25-30% of 22q11.2DS patients have hypoparathyroidism [46,47]. It leads to low calcium levels (hypocalcaemia) in the majority of these patients (90%) [47] and is possibly associated with autoimmune diseases [46]. Moreover, growth hormone deficiency has been observed in 22q11.2DS individuals [48,49] for which growth is known to be altered [50].

1.4.2. Neuropsychiatric/Neurological symptoms

Neuropsychiatric and cognitive impairments have a high incidence in 22q11.2DS patients. Learning disabilities were among the symptoms first described in the 22q11.2 deletion-associated VCFS syndrome [51]. It has since then been shown that around 55-60% of 22q11.2DS patients have a neuropsychiatric disorder [52–54]. A study of 32 patients with a confirmed 22q11.2 deletion (with ages ranging from 5 to 33 years) indicated that only 6% of them had a normal IQ and no psychiatric disorders. These disorders have varying incidence during the patients' lifetime and include notably psychotic disorders (schizophrenia in particular), mood and anxiety disorders, autism spectrum disorders (ASD), attention-deficit/hyperactivity disorders (ADHD) [55]. Many studies have investigated the cognitive and psychiatric changes in 22q11.2DS patients, as reviewed by Philip and Bassett (2011) [56], however most were of relatively small sample cohorts. The International 22q11.2 Brain Behaviour Consortium (22q11 IBBC) is an international effort to study 22q11.2DS genotypes, phenotypes and their relations at a large scale. The results from their published studies, including hundreds of 22q11.2DS patients, are discussed below.

1.4.2.1. Neuropsychiatric symptoms

The 22q11 IBBC published in 2014 a study that investigated psychiatric morbidity in 1402 patients with 22q11.2DS [57]. Participants (6-68 years, mean = 18.78, SD=10.66) were divided into 5 age groups: children (6–12 years), adolescents (13–17 years); emerging adults (18–25 years), young adults (26–35 years), and mature adults (≥36 years) (N = 456/346/323/150/127 respectively). The prevalence of each psychiatric disorder obtained in this study is described in Table 1-2. In children, the prevailing disorders observed were ADHD (37%), anxiety disorders (36%), disruptive disorders (14%) and ASD (13%). On the other hand, in adult there was a large incidence of schizophrenia spectrum disorders (23% to 42% depending on age group) which was mostly due the presence of schizophrenia (13% to 30%). These results were concordant with early studies showing a prevalence rate of psychotic disorder in 22q11.2DS adult patients of 30% and around 23-24% for schizophrenia [58,59]. The other most prevalent disorders in adults are anxiety disorders (24-28%), mood disorders (15-20%), ASD (16%) and ADHD (16%).

During the lifespan of these 22q11.2DS patients, there was a significant increase in the prevalence rate of schizophrenia spectrum disorder, panic disorders and major depressive disorder. Conversely, there was a decrease in the rates of most anxiety disorders [57]. There were significant comorbidity associations between anxiety disorders, mood disorders and schizophrenia. Gender was not shown to affect the prevalence of psychotic disorders but there was a higher frequency of disruptive disorders and ADHD in males and a higher frequency of mood and anxiety disorders in females (for adults).

1.4.2.2. Cognitive impairment and links to schizophrenia

It has been shown in the past that 22q11.2DS patients have on average a lower full-scale IQ, typically within the borderline intellectual functioning range (70-85) [52,60,61]. In the 22q11.2 IBBC cohort the average full-scale IQ was 71 points, with 46% of patients presenting intellectual disability scores (< 70) [57]. Moreover, cognitive impairments were associated with schizophrenic spectrum disorders. A longitudinal study by the same consortium (N = 829 patients) has reported that 22q11.2DS patients with an IQ lower at

Diagnosis	Children and Adolescents				Adults					
	Children (6–12 Years)		Adolescents (13–17 Years)		Emerging Adults (18–25 Years)		Young Adults (26–35 Years)		Mature Adults (≥36 Years)	
	N	%	N	%	N	%	N	%	N	%
Any schizophrenia spectrum disorder[a]	9/456	1.97	35/346	10.12	76/323	23.53	62/150	41.33	53/127	41.73
Schizophrenia	1/456	0.22	13/342	3.80	36/291	12.37	37/132	28.03	32/106	30.19
Schizoaffective disorder	0/456	0.00	3/342	0.88	5/291	1.72	10/132	7.58	4/106	3.77
Schizophreniform disorder	0/259	0.00	1/289	0.34	3/285	1.05	0/127	0.00	0/89	0.00
Brief psychotic disorder	0/259	0.00	4/289	1.38	1/288	0.35	1/131	0.76	0/106	0.00
Psychotic disorder not otherwise specified	8/456	1.75	14/346	4.05	29/323	8.98	14/150	9.33	17/126	13.49
Delusional disorder	0/456	0.00	0/346	0.00	2/323	0.62	0/150	0.00	0/127	0.00
Any anxiety disorder[b]	155/435	35.63	97/286	33.92	71/295	24.07	37/149	24.83	35/127	27.56
Separation anxiety disorder[c]	25/395	6.33	4/259	1.54	2/113	1.77	0/28	0.00	0/20	0.00
Specific phobia[d]	95/433	21.94	48/282	17.02	19/263	7.22	5/131	3.82	3/106	2.83
Social phobia[e]	45/435	10.34	28/286	9.79	14/295	4.75	4/149	2.68	1/127	0.79
Panic disorder[f]	4/333	1.20	2/231	0.87	17/270	6.30	12/137	8.76	17/118	14.41
Posttraumatic stress disorder	1/274	0.36	3/222	1.35	2/240	0.83	0/109	0.00	2/73	2.74
Obsessive-compulsive disorder	24/435	5.52	17/286	5.94	15/295	5.08	8/149	5.37	8/127	6.30
Generalized anxiety disorder	36/435	8.28	30/286	10.49	29/295	9.83	18/148	12.16	14/127	11.02
Anxiety disorder not otherwise specified	1/435	0.23	1/286	0.34	2/295	0.68	1/149	0.67	0/127	0.00
Any mood disorder	15/456	3.29	41/346	11.85	59/323	18.27	22/150	14.67	26/127	20.47
Major depressive disorder[g]	10/456	2.19	31/346	8.96	35/323	10.84	18/150	12.00	20/127	15.75
Dysthymia[h]	5/456	1.10	8/346	2.31	16/320	5.00	2/145	1.38	1/110	0.91
Bipolar disorder or (hypo)manic episode in children	0/318	0.00	2/317	0.32	6/320	1.88	3/150	2.00	5/127	3.94
Mood disorder not otherwise specified	0/456	0.00	4/346	1.16	7/323	2.17	0/150	0.00	2/127	1.57
Substance-related disorder (substance abuse and dependence)	0/300	0.00	1/221	0.45	7/278	2.52	9/142	6.34	5/110	4.55

	Children (6–12 Years)		Adolescents (13–17 Years)		Adults (≥18 years)	
	N	%	N	%	N	%
ADHD[i]	161/434	37.10	63/264	23.86	29/186	15.59
Autism spectrum disorders[j]	12/94	12.77	43/162	26.54	47/292	16.10
Any disruptive disorder[k]	57/400	14.25	25/229	10.92	9/127	7.09
Oppositional defiant disorder	57/400	14.25	25/229	14.79	7/115	6.09
Conduct disorder	0/316	0.00	0/180	0.00	2/138	1.45

[a] Significant increase with age (χ^2=214.70, df=4, p<0.001).
[b] Significant decrease with age (χ^2=15.49, df=4, p=0.004).
[c] Significant decrease with age (χ^2=13.67, df=4, p=0.008).
[d] Significant decrease with age (χ^2=57.13, df=4, p<0.001).
[e] Significant decrease with age (χ^2=24.57, df=4, p<0.001).
[f] Significant increase with age (χ^2=47.35, df=4, p<0.001).
[g] Significant increase with age (χ^2=37.88, df=4, p<0.001).
[h] Significant difference across age groups (χ^2=14.66, df=4, p=0.05).
[i] Significant difference between children and adolescents (χ^2=13.19, df=1, p<0.001) and between adolescents and adults (χ^2=4.59, df=1, p=0.04).
[j] Significant difference across age groups (χ^2=10.07, df=2, p=0.007).
[k] Nonsignificant difference across age groups (χ^2=5.06, df=2, p=0.08).

Table 1-2. Prevalence of psychiatric disorders (DSM-IV-TR) in five age groups of subjects with 22q11.2 DS. Figure from Schneider 2014 [57]**.** Reprinted with permission from The American Journal of Psychiatry, (Copyright©2014). American Psychiatric Association. All Rights Reserved.

first cognitive assessment had an increased rate of psychotic disorder in subsequent IQ with age, and this decline was significantly increased in the patients that developed a psychotic disorder compared to the ones that did not. However, this decrease in IQ in 22Q11.2DS patients was not observed in another recent study [62].

1.4.2.3. Early-onset Parkinson disease

Multiple clinical cases of 22q11.2 patients with early onset (< 45 years old) Parkinson's disease (PD) have been reported [63–66]. Butcher *et al*. reported in 2013 that the occurrence of early-onset PD was significantly increased in 22q11.2DS patients compared to the general population [67]. Similar results were shown in 2016 by Mok *et al.* who also demonstrated that in a large cohort of PD patients, the rate of previously undetected deletions at 22q11.2 was significantly greater than that seen in population controls, and that this was further enriched in patients with early onset PD [68].

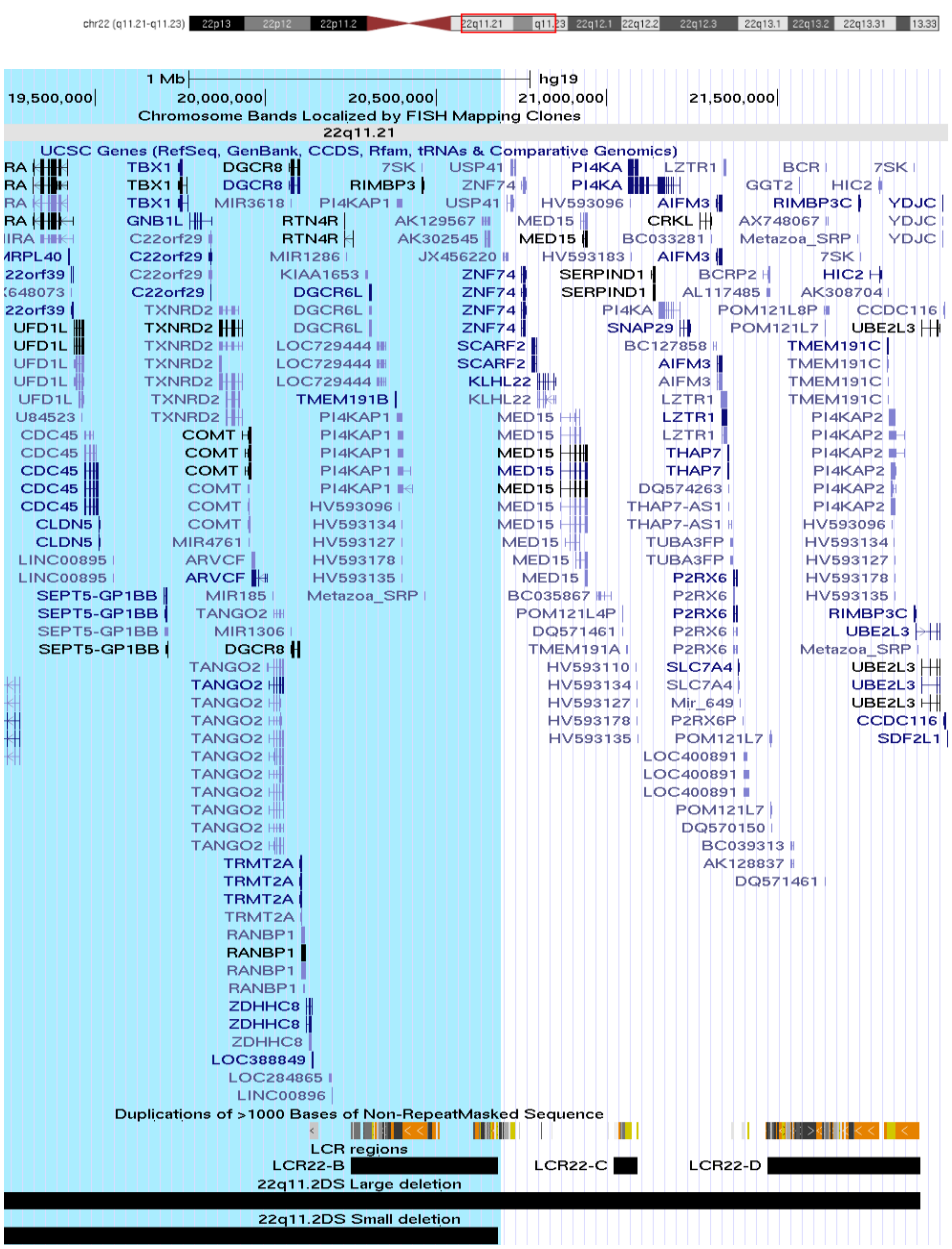
1.4.3. Reduced lifespan in 22q11.2DS patients

A retrospective study done by Repetto *et al.* in 2014 investigated fatality rates in 22q11.2DS from 0 to 52 years (median age of patients: 12 years old) [69]. They observed a median age of death of 3.4 months (ranging from 3 days to 32.4 years, with only two patients dying later than 2 years old). The study showed a significant association of mortality with the presence of congenital heart disorders, hypocalcaemia and airway malacia. A study including only adult 22q11.2DS patients showed that only 95%, 90% and 74% survived to age 30, 40 and 50 respectively, with a median age of death of 41.5 years (ranging from 18 to 69 in a study including only adults) [70]. Death rates in adults were not apparently associated with congenital heart disorders nor schizophrenia.

1.4.4. Deletion size and phenotypes

No study has identified a significant correlation between clinical and neuropsychiatric phenotypes and the size of the deletion (smaller A-B or larger A-D) [16,71,72]. However, due to the small percentage of 22q11.2DS patients with the A-B deletion (5-10%), statistical analysis is challenging. For instance, Michaelovsky *et al.* compared symptoms of 97 patients with the larger A-D deletion versus 8 patients with a smaller A-B deletion [71]. No patient with an A-B deletion presented hypocalcaemia, while around 30% of patients with the large A-D deletion had this symptom. Despite this, it was not possible to draw a definitive conclusion based on the relatively low number of patients. On the other hand,

A



B

Gene name
USP18
LOC100996415
LOC105379550
LINC01662
GGT3P
LOC102725072
LINC01663
DGCR6 *
PRODH *
DGCR5
DGCR9
DGCR10
DGCR2 *
DGCR11
DGCR14 *
TSSK2
GSC2
LINC01311
SLC25A1 *
CLTCL1 *
HIRA *
MRPL40 *
C22orf39 *
UFD1 *
CDC45 *
CLDN5 *
LINC00895
SEPT5 *
SEPT5-GP1BB
GP1BB *
TBX1 *
GNB1L *
RTL10 *
TXNRD2 *
COMT *
MIR4761
ARVCF *
TANGO2 *
MIR185 *
DGCR8 *
MIR3618
MIR1306 *
TRMT2A *
MIR6816
RANBP1 *
SNORA77B
ZDHHC8 *
CCDC188
LOC284865
LINC00896
RTN4R *
MIR1286 *
DGCR6L *
LOC101927859
FAM230A
TMEM191B
PI4KAP1
RIMBP3
LINC01660

Figure 1-3. Genes located within the 22q11.2 region. (**A**) **Transcripts included within the large and small 22q11.2 DS deletion.** The genes included within the small regions are highlighted in light blue. Source: UCSC Genome Browser [491]. (**B**) **List of all genes included within the LCR A to LCR B deletion.** Protein coding genes are indicated in bold. Brain expressed genes are noted with an asterisk * [74]. The genes are sorted by chromosomal location. Source: UCSC Genome Browser [491] - RefSeq genes data.

there is evidence indicating that patients with an atypical distal deletion that does not overlap with the A-B region might show different phenotypes. This type of deletion has not been detected in patients with typical VCFS phenotypes and/or conotruncal heart defects, but it has been reported in patients with milder and more atypical VCFS phenotypes [73].

1.5. 22q11.2 deletion: genes, synteny and animal models

1.5.1. Genes within the deletion

There are 92 genes spanned by the larger A-D 22q11.2 deletion, of which 53 are protein coding (UCSC RefSeq gene data) (Figure 1-3.A). 59 genes are included in the smaller ~1.5 Mb region that is flanked by LCR A and LCR B. This includes 32 protein coding genes (Figure 1-3.B), 27 of which (84%) have been shown to be expressed in the brain [74]. The other genes are microRNA genes (miR-4761, miR-185, miR-3618, miR-1306, miR-6816, miR-1286), non-coding RNA genes and pseudogenes.

1.5.2. Inter-species synteny

Genes within the 22q11.2 region appear to be well conserved between species. Comparative mapping of 46 protein coding genes contained within the human 22q11.2 A-D deletion has found 40 conserved homologues genes in *M.musculus*, 37 in *D.rerio*, 22 in *D.melanogaster* and 17 in *C.elegans* [74]. The mouse syntenic region is shown in Figure 1-4. This homology between species has allowed animal models to be created and these have facilitated studies of the impact of the heterozygous or homozygous deletions of 22q11.2 genes. Guna *et al.* have reviewed in 2015 all the available single-gene knock-out and knock-down models for 22q11.2 genes in these 4 animal species, as well as phenotypic observations [74]. In particular, 31 of the 40 mouse homologues of 22q11.2 deletion genes have available knock-out strains. Moreover, multigene deletions have also been generated, some of which include all 27 conserved protein coding genes homologous to the human LCRA to LCR D deletion genes (Figure 1-4).

1.5.3. 22q11.2DS mice model principal phenotypes

The deletion in mice of the region syntenic to the smaller A-B deletion (such as *Df1*, *LgDel* or *Df(16)A* models, Figure 1-4) has been shown to result in phenotypes that reproduce some of the clinical features observed in 22q11.2DS patients. In 1999 Lindsay

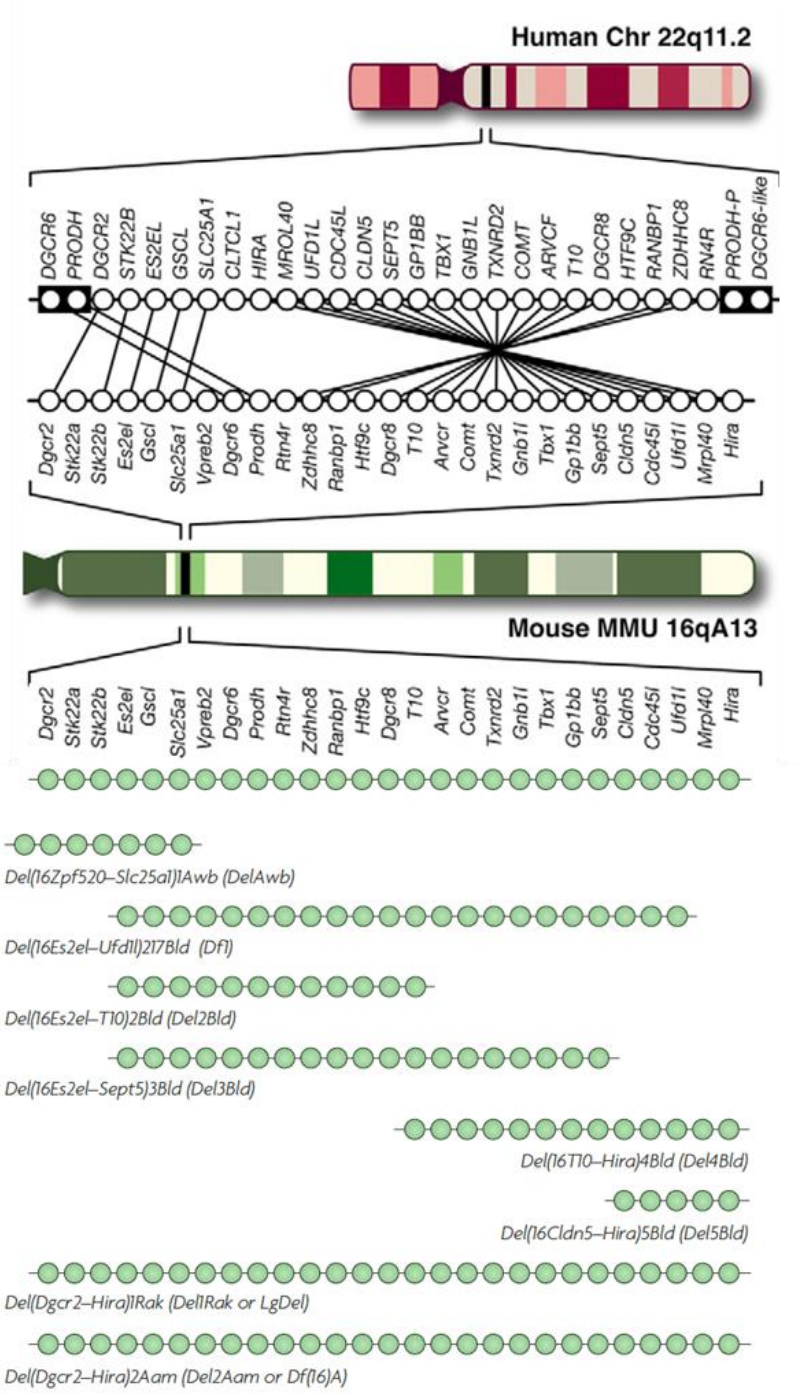


Figure 1-4. Human chromosome 22q11.2 region, syntenic mouse region and mouse models of 22q11.2 deletion. Each circle represents one gene. The lower part of the figure represents various multigene deletion models that have been characterized for neuronal and behavioural abnormalities labelled using their Mouse Genome Database (MGD) allele symbols and commonly used synonyms. Upper part: figure adapted from Arguello 2013 [493], used under CC-BY 3.0 licence; lower part: figure reprinted by permission from Springer Nature: Nature Reviews Neuroscience, Karayiorgou 2010 [494], copyright (2010).

et al. reported the first 22q11.2DS mouse model, *Df1/+* [75]. Mice from this strain exhibited cardiovascular defects similar to what is observed in 22q11.2Ds patients [35], such as

interrupted aortic archs or ventricular septal defects. Mersher *et al.* found similar cardiovascular defects in another model that included 5 more genes (*LgDel/+*, cf. Figure 1-4) [76]. Other observations in these mice included increased perinatal lethality and parathyroid hypoplasia, which are 22q11.2DS human phenotypes [46,47,69].

22q11.2DS mouse models also exhibited behavioural deficits. Paylor *et al.* reported that the *Df1/+* mice have an abnormal sensorimotor gating [77]. This has also been reported in the *Df(16)A*$^{+/-}$ 22q11.2DS mouse model [78]. The observed reduced pre-pulse inhibition has been considered to be an endophenotype of schizophrenia [79], and this reduction in pre-pulse inhibition has also been reported in children with 22q11.2DS [80]. Other reported behavioural changes for these 22q11.2DS mice includes hyperactivity, reduced freezing during contextual fear-conditioning tests and impaired spatial working memory-based learning (*Df(16A*$^{+/-}$*)* [78]. The deficits in spatial memory in adult 22q11.2DS mice have been linked to enhanced hippocampal long-term potentiation (LTP) [81–83].

1.6. 22q11.2 deletion: potential schizophrenia candidate genes

The observed rate of schizophrenia in 22q11.2DS adult patients (around 25-30%) [57–59] is more than 30 time higher than the lifetime morbid risk of schizophrenia in the general population of 0.72% (median lifetime morbid risk from a 2008 systematic review of schizophrenia prevalence rates, mean = 1.19%, sd. = 1.08% [84]). The 22q11.2 deletion is the highest known risk factor associated with schizophrenia and is present in around 0.3% of all schizophrenia patients [85–87]. For these reasons, this thesis is mainly focused on the study of schizophrenia in the context of 22q11.2DS. The similarity of neuropsychiatric symptoms in 22q11.2DS patients with both A-B and A-D deletions suggest that the genes relevant to these psychiatric features are included within the 1.5 Mb deletion spanning from LCR A to LCR B. The following sections briefly discuss the most relevant genes mapped to the A-B 22q11.2 deletion and their potential links to schizophrenia.

1.6.1. *COMT*

The *COMT* gene codes for the catechol-O-methyltransferase protein, an enzyme involved in the metabolic degradation of catecholamines such as the neurotransmitters dopamine and norepinephrine. The gene codes for two variants, one soluble (S-COMT) and one membrane bound (MB-COMT) [88]. A polymorphism located at position 108 and

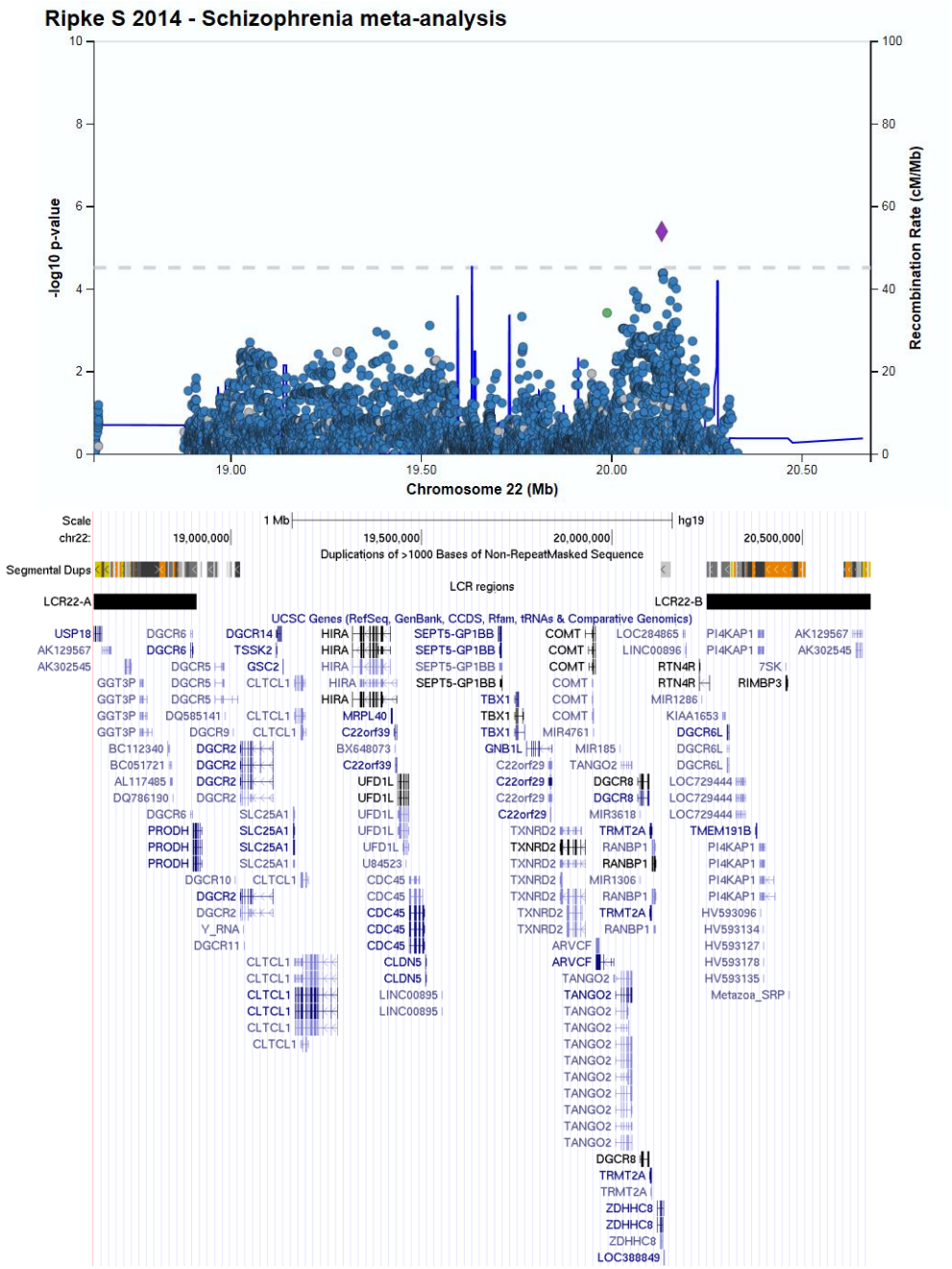


Figure 1-5. Schizophrenia GWAS results for the region 22q11.2 (LCR A to LCR B) (non-22q11.2DS specific analysis). Data from the summary statistic of the most powerful genome-wide association study (GWAS) to date (Ripke *et al.* 2014 [92]). Plot generated with LocusZoom.js (LocusZoom.org [495]) and UCSC genome browser [491]. No common polymorphism has p-value lower than the genome-wide significance threshold ($p = 5*10^{-8}$); thus no polymorphism within the small 22q11.2 deletion is significantly associated with schizophrenia in this study. The lowest P-value is associated with the *ZDHHC8* gene (purple diamond, $p = 3.98^{*}10^{-6}$).

158 of these two variants respectively (Val/Met, known as $Val_{158}Met$) have been shown in 1996 by Lachman *et al.* to be associated with the COMT enzyme activity [89]. It was later demonstrated that it was altering both the activity and quantity of the enzyme in post-mortem human brain [90]. Despite early studies showing an association of this

polymorphism with schizophrenia risk there is no consensus in more recent studies and its implication in schizophrenia is debated (as reviewed by Williams 2007 [91]). While no association between common polymorphisms at the COMT locus and schizophrenia has been detected in the largest schizophrenia genome-wide association study (2014, cf. Figure 1-5) [92], a meta-analysis performed by Gonzalez-Castro *et al.* in 2016 (including more than 15000 schizophrenia patients and 17000 controls) has shown an association between the Val_{158}Met allele and this psychiatric disorder [93]. Moreover, it has been shown that the DNA methylation levels (an epigenetic mark) of COMT isoforms promoters are altered in schizophrenia patients [94–96].

1.6.2. *PRODH*

The *PRODH* gene codes for the proline dehydrogenase 1, also named proline oxidase. Its homozygous deletion in mouse results in altered pre-pulse inhibition [97]. These animals have highly increased proline levels, which is consistent with the hyperprolinemia that have been observed in schizophrenia patients with PRODH mutations [98] but also in some 22q11.2DS patients [99]. In these 22q11.2DS patients with hyperprolinemia, high proline levels was associated with lower IQ [99]. Association studies investigating links between PRODH polymorphisms and schizophrenia have had mixed results, as reviewed in Willis 2008 [100].

1.6.3. *TBX1*

The *TBX1* gene (*T-box 1*) is a member of the transcription factors T-box gene family which are involved in development [101]. Multiple studies in 2001 have demonstrated that deletion of *Tbx1* in mice lead to cardiovascular defects in heterozygous mutants [76,102] as well as thymus and parathyroid hypoplasia and craniofacial defects in homozygous mutants [103]. *TBX1* has therefore been proposed to be one of main genetic drivers related to the abnormal development and congenital disorders observed in 22q11.1DS patients. It has been suggested that these observed phenotypes are due to an abnormal migration and distribution of the neural crest cells [104] which are involved in the development of these tissues [105–107]. The heterozygous deletion of *Tbx1* has been shown to affect sensorimotor gating in mice (reduced pre-pulse inhibition) [77]. Moreover $Tbx1^{+/-}$ also present behavioural deficits related to autism spectrum disorders such impaired working memory, ultrasonic vocalisation and social interaction [108]. Finally, mesoderm-specific

heterozygous expression of *Tbx1* also results in abnormal cortical development [109]. Clinical case studies have reported patients with specific mutations of *TBX1* but without the 22q11.2 deletion that presented similar clinical features than 22q11.2DS patients, such as the characteristic facial features, congenital heart disorders and neuropsychiatric disorders (ASD, mental retardation or behavioural issues) [77,110,111].

1.6.4. *GNB1L*

The heterozygous knock-out of the *Gnb1l* gene (*guanine nucleotide-binding protein, beta-1-like*) in mouse has shown reduced pre-pulse inhibition [77]. Moreover, while single nucleotide polymorphisms (SNPs) within this gene have been associated with increased schizophrenia risk in association studies [112,113], the results have not been replicated in the most powerful genome-wide association study to date [92] (Figure 1-5). Nevertheless, it has been shown that some *GNB1L* SNPs alleles are expression quantitative trait loci (eQTLs)[113,114], which means that they are correlated with, and could possibly have an effect on gene expression. This is of particular interest as *GNB1L* mRNA and protein have been shown to be reduced in brain tissue of schizophrenia patients, and that treatment with haloperidol, an antipsychotic drug, in mouse lead to increase *Gnb1l* expression [115].

1.6.5. *ZDHHC8*

Early studies have associated SNPs within the *ZDHHC8* gene (*zinc finger DHHC domain-containing protein 8*) with schizophrenia risk [116–118]. However, more recent association studies (trio studies and genome-wide association studies) in larger samples have failed to replicate this result [92,119] (Figure 1-5). Nonetheless, *ZDHHC8* is an interesting candidate to study 22q11.2DS neuropsychiatric disorders. This gene codes for a brain-expressed palmitoyltransferase [118]. Palmitoylation is a type of lipid modification that has been shown to be important for neuronal development and synaptic plasticity, targeting proteins such as the post-synaptic protein PSD95 and paralemmin 1 protein which are essential for dendritic spine formation (as reviewed in Fukata 2010 [120]). Mice knocked-out of this gene have been shown to have behavioural effects such as decreased pre-pulse inhibition [118] and impaired spatial working memory [121]. Primary neuronal culture from *Zdhhc8*$^{+/-}$ mice has shown a reduced dendritic complexity and spine density that was also observed for a mouse model of 22q11.2DS [122]. Interestingly, transfecting ZDHHC8 back into neurons of this mouse model resulted in the abnormal spine density

phenotype being rescued. Studies suggest that the observed dendritic and as well as axonal deficits in *Zdhhc8*$^{+/-}$ mice are due to abnormal palmitoylation of PSD95 and particularly CDC42, a protein involved in the Akt/Gsk3β pathway [121–123].

1.6.6. DGCR2

Polymorphisms within the *DGCR2* gene (*DiGeorge syndrome critical region Gene 2*) have been associated with schizophrenia and a reduction of *DGCR2* mRNA levels [124]. While this association has not been replicated in the most powerful schizophrenia GWAS [92] (Figure 1-5), other lines of evidence support potential links between *DGCR2* and schizophrenia. Firstly, exome sequencing identified a potentially disruptive *de novo* mutation in this gene in a schizophrenia patient [125], while secondly, its expression has been found to be greater in the dorsolateral prefrontal cortex of schizophrenic patients compared to individuals without neuropsychiatric disorders [124]. Moreover, treatment with antipsychotics such as clozapine lead to higher levels of *DGCR2* mRNA in rat prefrontal cortex [124].

1.6.7. *DGCR8*

DGCR8 (coded by *DiGeorge syndrome critical region Gene 8*) is a protein part of the microprocessor complex involved in the canonical microRNA biogenesis pathway [126]. MicroRNAs are small RNAs molecules (~22 bp) that have been shown to have a key role in the posttranscriptional regulation of gene expression [127]. *Dgcr8* deletion leads to reduced microRNA levels in mouse embryonic stem cells [128]. A reduction of microRNA levels was similarly found in the brain of *Dgcr8*$^{+/-}$ mice and the *Df(16)A*$^{+/-}$ 22q11.2 mice [78]. This dysregulation of microRNA levels was also found in the blood of 22q11.2DS patients [129]. Analysis of post-mortem brain from schizophrenia patients has shown that a subset of microRNAs had significantly altered levels, which were reduced in most cases [130–133]. Many microRNAs have been linked to schizophrenia, as reviewed in Beveridge and Cairns (2012) [134]. The potential role of DGCR8 in the aetiology of schizophrenia in the context of 22q11.2DS is further discussed later in this introduction (section 1.7.3.4 p.24) as well as in Chapter 4 (p.81) and Chapter 5 (p.106).

1.7. Phenotypic variability and neuropsychiatric disorders in 22q11.2DS.

22q11.2DS diagnosis is defined by the presence of the hemizygous deletion at chromosome 22q11.2. However, despite the primary mutation being common between all patients there is an important heterogeneity of symptoms between these individuals with the deletion. Some 22q11.2DS patients are relatively asymptomatic and the presence of the deletion is only discovered due to its transmission to a child in familial cases [9]. While the incidence of congenital disorders such as cardiovascular and craniofacial defects tend to be high [35,36,38], the incidence of neuropsychiatric disorders is more variable (Table 1-2). The following sections will review possible molecular mechanisms that could explain the phenotypic variability observed in 22q11.2DS patients.

1.7.1. Deletion size and breakpoint heterogeneity

The different size and breakpoints (proximal, central, distal) of the 22q11.2DS deletion could lead to a phenotypic variability between patients by encompassing different set of genes. Lopez-Rivera *et al.* (2017) have screened 2080 patients with kidney and urinary tract congenital defects for the 22q11.2 deletion and have detected 14 patients with a deletion at chromosome 22q11.2 [135]. Three types of deletions were detected in these patients: large proximal deletion (LCR A to LCR D), and two types of central deletions (LCR B to LCR D and LCR C to LCR D, cf. Figure 1-1.B). They all had in common the region going from LCR C to LCR D; and further work associated these congenital defects to the deletion of the *CRKL* gene included in this region [135]. In consequence, patients with either the typical small (LCR A to B) or large (LCR A to D) proximal 22q11.2 deletion are likely to show different renal phenotypes due to the deletion (or not) of this gene.

No differences in neuropsychiatric symptoms have been observed between patients with either A-B or A-D deletions [16,71,72]. This indicates that the main genetic driver of these features is included within the A-B deletion, which is nested within the larger A-D deletion.

However, while all patients with these deletions will share a common set of deleted genes, the exact size of the deletion may vary. Non-allelic homologous recombination (NAHR) events leading to the deletion occurs at the LCR regions (Figure 1-1.A), but the exact breakpoint within these LCRs has been shown to vary, resulting in different deletion

sizes. Weksberg *et al.* used qPCR analysis to identify that some patients reported to have an A-D 22q11.2 deletion still had both chromosomal copies of the region between *PRODH* and *DGCR2*, while this region was included within the deletion of the majority of patients [136]. The variability in deletion breakpoints between 22q11.2DS patients was also confirmed using a more precise method (microarray comparative genomic hybridisation) [32,137]. In a set of 20 patients, the size of the deletion ranged from 2.49 Mb to 3.09 Mb due to different breakpoints; for instance, 6 different breakpoints were detected within LCR-A [32]. As a consequence, genes such as *DGCR6* or *PRODH* are not necessarily included within the deletion for all patients and it might lead to inter-patient symptomatic variability. In particular, *PRODH* has been shown to be potentially associated with hyperprolinemia, IQ and schizophrenia [97–100].

1.7.2. Variability of expression of dosage-sensitive genes

The deletion of 22q11.2 genes will, theoretically, lead to a reduction in gene expression by half. This expression could however differ from this expected result due to different mechanisms of post-transcriptional regulation and homeostasis. Some genes, called dosage-sensitive genes, are particularly affected by such changes in expression and are more likely than others to lead to phenotypic changes [138]. These phenotypes can occur if the amount of protein coded by a dosage-sensitive gene is below or above a certain threshold. The heterozygous knock-out of single genes within 22q11.2 syntenic region in mouse models suggest that certain genes within the deletion are dosage-sensitive (cf. section 1.6 page 14). Moreover, chromosomal duplication of the 22q11.2 region (cf. Figure 1-1.B) has been reported to be associated with a decreased risk for schizophrenia [139]. This evidence reinforces the hypothesis that genes within the 22q11.2 region that are potentially involved in neuropsychiatric disease are dosage sensitive.

Gene expression is variable between individuals, and will depends on other factors, such as inherited variation [140] as well as environmental signals [141]. These different effects on gene expression will be added to the changes due to the hemizygous deletion of 22q11.2 genes. This will lead to a heterogeneity on the expression of these genes between individuals, and for dosage-sensitive genes some 22q11.2DS patients might have enough of the protein to not develop a phenotype while others will have the opposite effect. This may result in an apparent symptomatic heterogeneity.

1.7.3. The multiple-hit model

The two-hit -or multiple hit- hypothesis was first proposed and later formulated by Nordling (1953) and Knudson (1971) respectively in the context of cancer [142,143]. It proposes that cancer aetiology can be due to the presence of an accumulation of two or more mutations/genetic variants. It has since been applied to other disorders such as schizophrenia [144]. It commonly refers to the accumulation of multiple risk factors for a disorder, that can be either genetic or environmental. For instance, Girirajan *et al.* have reported in 2010 evidences for such model for the developmental delay-associated deletion 16p12.1 [145]. Patients with this deletion were more likely to have a second large CNV (> 500kb) or chromosomal abnormality than controls. Moreover, patients with the 16p12.1 deletion and a second "hit" had different or more severe phenotypes than the patients with only the 16p12.1 deletion.

The following sections review different mechanisms that could apply to explain the development of psychotic disorders in 22q11.2DS symptoms, in the hypothesis of a multiple-hit model.

1.7.3.1. Allelic variation within the intact 22q11.2 chromosome

The deletion at chromosome 22q11.2 is hemizygous; one copy of all deleted genes is still present in 22q11.2DS patients. Mutations within this region that are present on the non-deleted allele could potentially affect the expression of these genes and/or the structure or function of the proteins coded by some of these genes. For instance, Hoogendoorm *et al.* have shown in 2004 *in vitro* (with reporter gene vectors) that different promoter haplotypes could lead to changes in expression of more than 1.5 fold in genes such as PRODH or DGCR14 [146]. Association studies in non-deleted individuals have been performed to try and determine if variants such as SNPs in genes of the 22q11.2 region could be linked with an increased risk of schizophrenia (for instance in Liu 2002 [116]). Some positive association with schizophrenia have been reported (cf. section 1.6 page 14) but no common variants within the 22q11.2 deletion have been reported in the most powerful to date genome-wide association study [92] (Figure 1-5). However, variants within this region could have a different effect in 22q11.2DS patients than in the general population due to the loss of one copy of the genes.

To investigate this issue Hestand *et al.* (2016) curated around 12,000 hemizygous variant positions after performing targeted resequencing of the deletion region in 127 22q11.2DS patients [147]. They identified in the coding regions of 22q11.2 genes 95 non-synonymous variants, as well as three stop codon gains and two frameshift insertions. No significant association with schizophrenia has been found in a genetic association study testing single nucleotide variants, small insertions/deletions mutations and copy number variants within the non-deleted 22q11.2 chromosome [148]. However, in comparing 40 22q11.2DS patients with psychotic disorders and 48 22q11.2DS patients without, this study was probably underpowered to detect such effects and the analysis will need to be repeated with larger sample size cohorts.

Variation within the intact alleles could however be associated with other neuropsychiatric disorders and/or affects schizophrenia endophenotypes. For instance, the *COMT* Val_{158}Met polymorphism has been associated with multiple neuropsychiatric features. The *COMT* Met_{158} allele has been found to be associated with both ADHD and OCD in 22q11.2DS patients [149]. Studies have also studied the impact of this Met_{158} allele in cognition in 22q11.2DS patients, with however mixed results [150–153].

1.7.3.2. Importance of the genetic background

Genetic association studies have identified that the genetic risk for schizophrenia involves multiple classes of mutations, as reviewed by Rees *et al.* (2015) [154]. Currently, the analyses have focussed on two main classes of mutations, copy number variants (CNVs) and single-nucleotides variants (SNV), both of which can be either inherited or *de novo* mutations. The allelic frequency of these risk variants can vary from being common (frequency greater than 1% in the general population) to rare (<1%). In schizophrenia there is evidence for risk variants covering the full frequency spectrum [154]. This includes predominantly rare CNVs such as 22q11.2DS, rare SNVs and insertion/deletion (indel) mutations as well as common SNVs (also named common single-nucleotide polymorphisms, or SNPs). In 22q11.2DS patients, such genetic variants that are located outside the deleted region at 22q11.2 could contribute additively to the total risk for schizophrenia (or other clinical features).

This could be for instance one or more other CNVs. It has been estimated that around 5 to 10% of the genome contributes to these mutations; however some CNVs affect

possible non-essential genes and do not seem to have a phenotypic effect [155]. In a cohort of 20 22q11.2DS patients, Bittel *et al.* detected the presence of 254 CNVs ranging from 400 bp to 2.4 Mb (94 duplications, 160 deletions) [32]. The frequency and size of these CNVs appeared to be similar than in the general paediatric population. At least 11 CNVs have been previously implicated with an increased risk for schizophrenia (as reviewed in Kirov 2015) [156]. Presence of one of these CNVs could thus lead to further risk in 22q11.2DS patients, and potentially other CNVs could increase this risk interactively too due to the presence of the 22q11.2 deletion.

In 2013 Williams *et al.* tested the two-hit hypothesis of schizophrenia in the context of 22q11.2DS by specifically investigating CNVs presence in 48 22q11.2DS patients with (N=23) or without (N=25) schizophrenia [157]. This study revealed that 22q11.2DS patients with psychosis had on average significantly larger CNVs than those without. A similar study was performed in 2017 by the International 22q11.2 Brain Behaviour Consortium (22q11 IBBC) with increased sample size (N=329 adults (>25 years old) with 22q11.2DS, 158 with any schizophrenia spectrum disorder and 171 non-psychotic patients) [158]. It did not report significant between-group differences in the number or size of CNVs, but the number of genes included within the rare exonic deletions was significantly greater in the 22q11.2DS patients with schizophrenia. Gene-set enrichment analysis for genes within the CNVs (duplication or deletions) notably showed an enrichment in "Nervous system phenotype" genes among the duplicated genes.

Interestingly, another study comparing 22q11.2DS patients with or without congenital heart disorders (CHD) revealed a significant enrichment of a specific CNV (a duplication at 12p13.31) in the CHD group [159]. This further supported the two-hit model hypothesis in the development of specific symptoms for 22q11.2DS patients.

Merico *et al.* (2015) assessed the presence of multiple classes of variants (common polymorphisms, structural variants and CNVs) in whole-genome sequencing data from 22q11.2DS patients with or without schizophrenia [160]. They identified an increased number of rare deleterious variants that were enriched in neurofunctional protein-coding genes in 22q11.2DS patients with schizophrenia. However, this nominally significant difference did not survive multiple-testing correction and the study was underpowered (N=6 for 22q11.2DS patients with schizophrenia, N=3 without). A 2016 study by Tansey *et al*. reported that schizophrenia patients with a known pathogenic CNV (such as 22q11.2

deletion) had an increased burden of common risk alleles compared to the general population [161]. While this study was underpowered to detect differences between schizophrenic and non-schizophrenic patients for specific CNVs, it supports a multiple-hit model in which schizophrenia risk in pathogenic CNV carriers is also due to the presence of other risk variants. The multiple-hit CNV/common variants hypothesis in the context of 22q11.2DS is discussed further and explored in Chapter 2 (p.31).

1.7.3.3. Gene-environment interaction

Multiple clinical case studies have reported monozygotic twins with a 22q11.2 deletion but with discordant phenotypes [162–165].While it is unclear if the size of the deletion was identical in these twins due to the limited resolution of the method of detection (FISH) used in most of these studies, one study used a microarray-based detection and did report that the deletion breakpoints between two monozygotic twins were discordant [165]. Only one gene (*GGT3P*), included within the LCR-A region, differed between the deletions in the two twins, with the patient having the *GGT3P*$^{+/-}$ genotype presenting with more severe features of 22q11.2DS and dying prenatally (6 weeks). The inclusion or not of *GGT3P* in the deletion has not been associated to date with a different phenotype in other 22q11.2DS patients. The phenotypes assessed in these clinical case studies were non-psychiatric. The presence of different phenotypes between monozygotic twins could be due to the presence of postzygotic *de novo* mutations (such as other CNVs) but also due to environmental factors.

It has been suggested that genetic and environmental interactions could increase schizophrenia risk, as reviewed by Brown (2011) [166]. Possible environmental risk factors include for instance maternal infection during pregnancy, prenatal nutrients deficiencies, cannabis use and childhood trauma [166]. Environmental factors have been shown to alter epigenetic mechanisms such as DNA methylation and result in changes in gene expression [167]. Such events could occur in 22q11.2DS patients and partly explain the observed heterogeneity of symptoms.

1.7.3.4. Epistatic interactions

Epistasis is a term related to the interactions between genes. It has been used to define different ideas, as reviewed by Phillips (2008) [168]. In the context of complex traits,

epistatic interactions are commonly defined as "Statistical interactions between loci in their effect on a trait such that the impact of a particular single-locus genotype depends on the genotype at other loci" [169]. This can be interpreted as a deviation from the expected outcome from the cumulative effect of two (or more) alleles or mutations (*i.e.* the cumulative effect of allele A and allele B is different from the sum of the effect of allele A and the effect of allele B). Epistatic interactions can for instance occur between genes with a redundant function or even between genes involved in the same regulatory pathways. The mechanisms of epistatic interactions are reviewed in Lehner 2011 [170].

Epistatic interactions affecting phenotypes in 22q11.2DS patients could occur between genes located within the deletion. Paternili *et al.* described in 2005 a murine model of PRODH deficiency with increased expression of COMT mRNA and protein levels [171]. Moreover, COMT inhibition led to behavioural changes in PRODH mutants (compared to wild-type mice with COMT inhibition), suggesting the presence of an epistatic interaction between these two 22q11.2 genes. It has been shown that 22q11.2DS patients that have both the COMT Met_{158} allele and hyperprolinemia have an increased risk for psychosis (but not patients with only one of the two) [99]. Interaction between the Met_{158} allele and high proline levels have also been associated with ASD symptom severity in 22q11.2DS patients [172].

Epistatic interactions could also occur between genes within the deletion and genes located elsewhere in the genome. However, analysis of epistasis in complex traits such as schizophrenia is statistically and computationally challenging and will likely require large sample size to get significant results [169].

DGCR8 is one particularly interesting candidate for epistatic interactions due to the function of its protein in the microRNA biogenesis pathway [173]. MicroRNAs are regulators of gene expression and can be regarded as a "buffering" system that reduce noise in gene expression [174]. By being involved in homeostatic mechanisms, it can potentially reduce fluctuations in mRNA levels due to a particular stress or mutations. This also potentially allows the accumulation of cryptic mutations that are not expressed. However, in 22q11.2DS where DGCR8 is deleted, the effects of these previously silenced mutations could be expressed (Figure 1-6).

Grice *et al.* (2015) have investigated the effect of deleting one copy 22q11.2 genes in *Drosophila* (genes orthologues of *DGCR8*, *HIRA*, *SLC25A1*, *SEPT5*, *TBX1*, *GLBN1*,

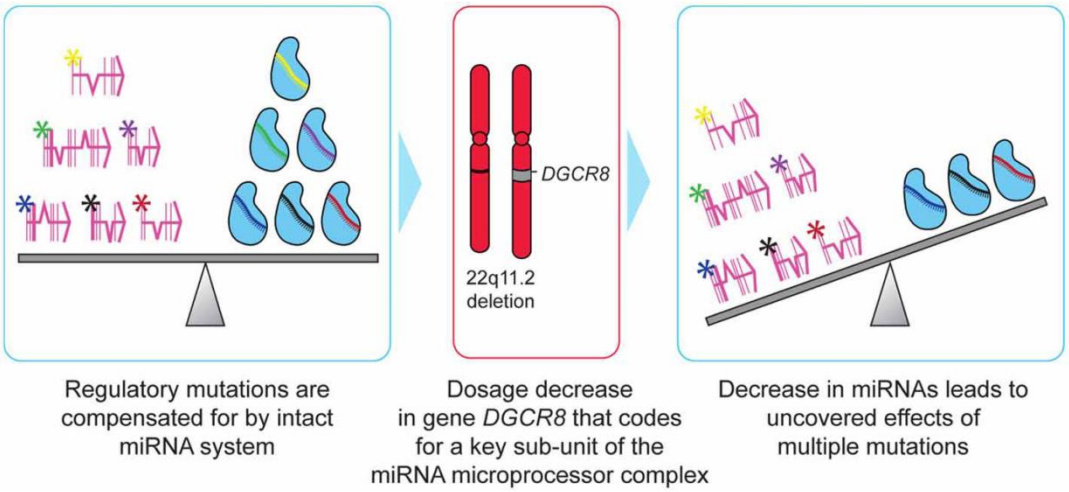


Figure 1-6. DGCR8-mediated epistatic interactions in 22q11.2DS. Figure from Brzustowicz 2012 [480], used under CC BY 3.0 license.

ZDHHC8) [175]. They examined the number of neuromuscular junctions (NMJ) bouton number and circadian rhythms of these mutant flies, and did not observed any significant difference. However, they did observed differences when they examined double-knockouts of *DGCR8* (*pasha* in drosophila) as well as *TBX1* (*org-1*) or *DGCR8* and *SEPT5* (*sep4*), which resulted in reduced NMJ bouton number and altered circadian rhythms. This result supports the hypothesis of DGCR8-mediated epistatic interactions within the 22q11.2 deletion, and could potentially be extrapolated to interactions with genes located outside this region.

The hemizygous deletion of DGCR8 could also interact with other mutations affecting microRNAs, further affecting their expression levels. For instance, microRNA-185 (miR-185) is among the genes deleted in 22q11.2DS (cf. Figure 1-3) but is also regulated by DGCR8, thus resulting in more than 50% reduction of the levels of this microRNA in the brain of 22q11.2DS mice (*Df(16)A*$^{+/-}$ strain) [176]. Moreover, reduction of miR-185 has been shown to cause neuromorphological changes (such as reduced dendritic complexity) [176]. MiR-185 has also been shown to target the DNA methyltransferase protein DNMT1 [177,178]. DNA methylation is an epigenetic mark known to affect gene expression and to be impacted by gene variation [179], potentially adding another layer of complexity in the possible epistatic interactions between the 22q11.2 deletion and the rest of the genome. A detailed description of the relationship between DNA methylation and 22q11.2DS are further discussed in Chapter 3 (p.49).

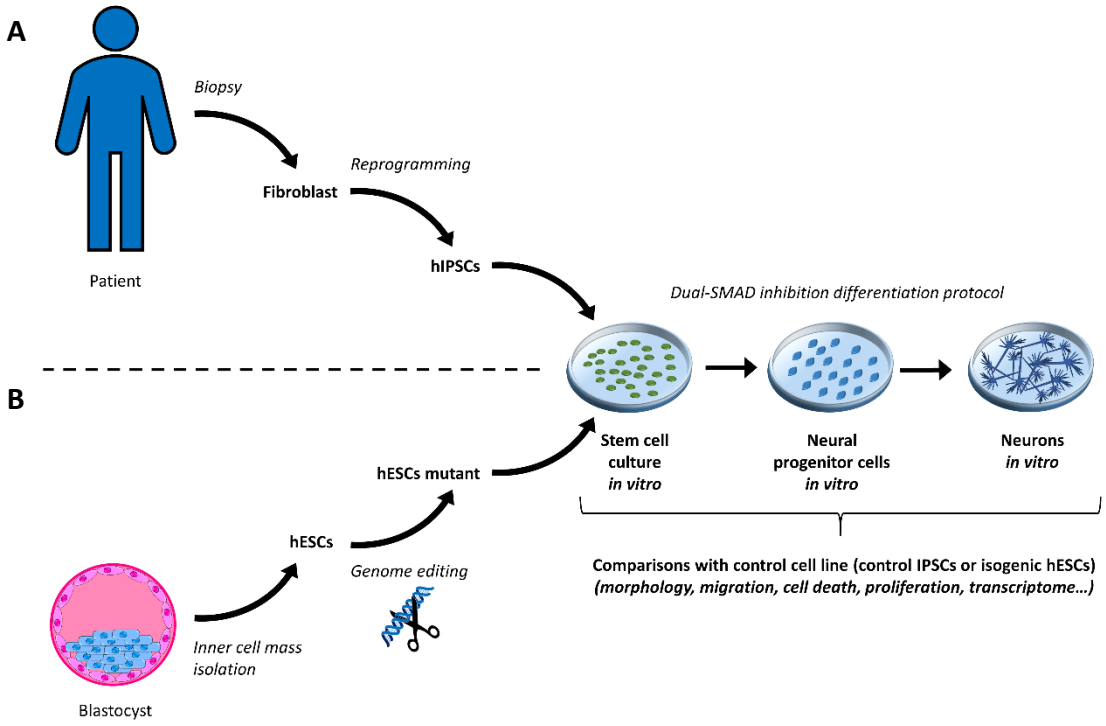


Figure 1-7. Stem-cell derived models of schizophrenia. (**A**) **Generation of human induced pluripotent stem cells (hiPSCs)**: fibroblasts are derived from a biopsy (*e.g.* skin, hair) from a schizophrenia patient and then reprogrammed into hiPSCs. These hiPSCs will have the same genetic background as the patients, including schizophrenia-associated variants (such as 22q11.2 deletion). (**B**) **Generation of human embryonic stem cells (hESCs) mutants**. hESCs are derived from the inner cell mass of a human blastocyst; then a potential schizophrenia gene candidate can be modified with genome editing (*e.g.* CRISPR/Cas9, TALEN) to obtain a mutant line to test (such as a 22q11.2 gene knock-out). Pluripotent stem cells (hiPSCs/hESCs) can then be differentiated into neural progenitor cells and neurons to test a phenotype of interest.

Other mutations affecting microRNAs could act synergistically with the dysregulation of microRNA regulation due to DGCR8 haploinsufficiency. Analysis of the gene content of CNVs outside of 22q11.2 in 22q11.2DS patients has shown that some of them are overlapping a microRNA gene [137,158]. While there was not an enrichment of "miRNA-CNVs" in 22q11.2DS patients with schizophrenia (compared to 22q11.2DS patients without) [158], it has been shown that miRNA-CNVs are involved in biological pathways relevant to 22q11.2DS symptoms [137].

1.8. Modelling schizophrenia with pluripotent stem cells

Murray and Lewis (1987) formulated the hypothesis that schizophrenia is a neurodevelopmental disorder [180]. Nowadays, this hypothesis is the most commonly accepted for schizophrenia aetiology, and states that neurodevelopmental deficits leading to schizophrenia start developing *in utero* and continue to develop until adolescence /

early adulthood. This is supported by multiple lines of evidences, from longitudinal studies of neuropsychiatric disorders to neuroimaging studies, as reviewed in Rapoport *et al.* (2012) [181]. Due to the limited availability of human brain tissue during development for research studies, *in vitro* neural differentiation of human pluripotent stem cells to model schizophrenia is a valid alternative. For this, two types of pluripotent stem cells are particularly interesting. The first ones are human induced pluripotent stem cells (hiPSCs). Tahakashi *et al.* demonstrated in 2007 that it is possible to reprogram differentiated cells (adult human fibroblasts) into hiPSCs by expressing only 4 factors: Oct3/4, Sox2, Klf4 and c-Myc (commonly called Yamanaka factors) [182]. These cells will have the same genetic background than the original patients, including disease-associated mutations (Figure 1-7.A). Another type of pluripotent stem cells that can be used to study schizophrenia is human embryonic stem cells (hESCs). Thomson *et al.* (1998) demonstrated that it is possible to isolated these cells from the inner cell mass of human blastocysts [183]. It is then possible to edit the genome of these cells using technologies such as TALEN or CRISPR/Cas9 to introduce specific disease-associated mutations (as reviewed by Musunuru 2013 [184]) (Figure 1-7.B). These two type of stem cells (hiPSCs/hESCs) can then be differentiated into cells from the neural lineage such as neural progenitor cells and cortical neurons [185,186]. This allows the study of potential defects occurring during the formation and maturation of neurons due to schizophrenia-specific mutations, such as the 22q11.2 deletion. For instance, by generating hiPSCs from patients with Rett syndrome, Marchetto *et al.* generated an autism spectrum disorder cell model. This allowed them to demonstrate that neurons derived *in vitro* from these cells had both morphological and physiological deficits compared to neurons derived from control hiPSCs [187].

Multiple stem cell models of schizophrenia have already been studied. Wen *et al.* have generated hiPSCs from patients with mutations in the schizophrenia-associated gene DISC1 [188], allowing them to observe that after neural differentiation these cells exhibited synaptic deficits and had an altered transcriptome [189]. *DISC1* disruption by genome-editing in control hiPSCs has also been shown to affect cell fate during neuronal differentiation [190]. Neural progenitor cells (NPC) are neurons' precursor cells and as early defects observed at this stage are likely to have repercussions later during development *in vivo*, they are generally considered to be a good model to study neurodevelopmental defects. It has been shown that NPCs generated from hiPSCs of schizophrenia patients

exhibit multiple deficits such as reduced cell migration as well as transcriptomic and proteomic changes [191,192]. The use of neurons and NPCs derived CRISPR/Cas9-edited hESCs and patient-specific hiPSCs in the context of 22q11.2DS and schizophrenia is further discussed in Chapter 4 (p.81) and Chapter 5 (p.106).

1.9. Aims of the thesis

The main aims of the thesis are:

- To investigate the genetic and epigenetic background of 22q11.2DS patients and their possible links to schizophrenia.
- To explore the role of DGCR8 during cortical neuron differentiation in human cell lines and its effect on gene expression.

This thesis is divided into 4 results chapters with the following objectives:

- In the Chapter 2 (p.31), I investigated the burden of common polymorphisms (SNPs) associated with schizophrenia in 22q11.2DS patients. For this I used a polygenic risk score analysis approach [193,194] that allowed me to determine if the 22q11.2DS patients with schizophrenia had more schizophrenia-associated SNPs than 22q11.2DS patients with no psychotic disorders or individuals with no pathogenic CNV and no psychotic disorders. It permitted to assess if these variants are playing a role in the aetiology of this neuropsychiatric disorder in 22q11.2DS patients.

- In the Chapter 3 (p.49), I tried to determine changes in DNA methylation due to the 22q11.2 deletion. I used a microarray-based method to compare the methylome of individuals with or without the deletion to see if there was significant changes due to these CNVs, and if these changes were affecting known schizophrenia genes or pathways.

- In the Chapter 4 (p.81), I used a recent genome editing technology (CRISPR/Cas9) to generate DGCR8 knock-out human embryonic stem cells (hESCs). The goal of this chapter was to obtain haploinsufficient DGCR8$^{+/-}$ hESCs line that can be differentiated into neuronal precursor cells to study the impact of DGCR8 depletion during neurodifferentiation of human cell *in vitro*. This chapter focussed on the generation of these cell lines and on providing evidence of haploinsufficiency.

- In the Chapter 5 (p.106), I investigated the mRNA transcriptome of DGCR8$^{+/-}$ neural progenitor cells to determine the impact of the haploinsufficiency on gene expression during neurodifferentiation. I then compared these results to the transcriptome changes due to the A-D deletion at chromosome 22q11.2 in human induced pluripotent stem cell-derived neural progenitor cells to try and determine common affected genes and pathways.

2. Schizophrenia polygenic risk score analysis in 22q11.2 deletion syndrome patients

2.1. Introduction

2.1.1. Common polygenic variation and schizophrenia

Schizophrenia has a strong genetic component leading to a high heritability in liability, as it has been shown by the high concordance rates in monozygotic twins (41-65%) [195]. Sullivan *et al.* (2003) estimated that heritability to be around 80% [196]. Schizophrenia heritability has been shown to follow a polygenic model rather than a single-locus model [197,198]. The genetic risk is a combination of alleles present in the population with high to low frequency, called common and rare variants/alleles respectively, the large majority of which have only a small effect on schizophrenia risk [154]. Genome wide-association studies (GWAS) have been able to identify common variants that are associated with schizophrenia [199]. The largest study to date was conducted by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) and consisted of around 150,000 samples (~1:3 cases/controls ratio) [92]. It reported 108 independent genetic loci to be associated with schizophrenia at a level that exceeded the threshold for genome wide significance ($p \leq 5*10^{-8}$, as represented by the red line in Figure 2-1.A) [92]. Larger sample sizes and advancement of analysis methods will likely lead to the discovery of more common variants associated with this disorder.

2.1.2. Polygenic risk score

Based on the multifactorial threshold model [200], an individual can be considered to develop schizophrenia when their total liability, which is composed of the cumulative effects of all genetic risk alleles carried (total genetic risk) as well as their exposure to environmental risk factors, is greater than a liability threshold (Figure 2-1.D). It is estimated that single-nucleotide polymorphisms (SNPs) cumulatively contribute around 23% of the variance in liability in schizophrenia, mostly due to common variants [201]. The aggregate risk to a disorder contributed by such common genetic variants can be calculated with a polygenic risk score (PRS) [193]. It can be calculated with GWAS summary association statistics, that contains information about each SNP tested in GWAS, such as per-allele effect size and association p-value. The effect size, in the case of a binary phenotype such as schizophrenia is defined as a log odd-ratio and represents

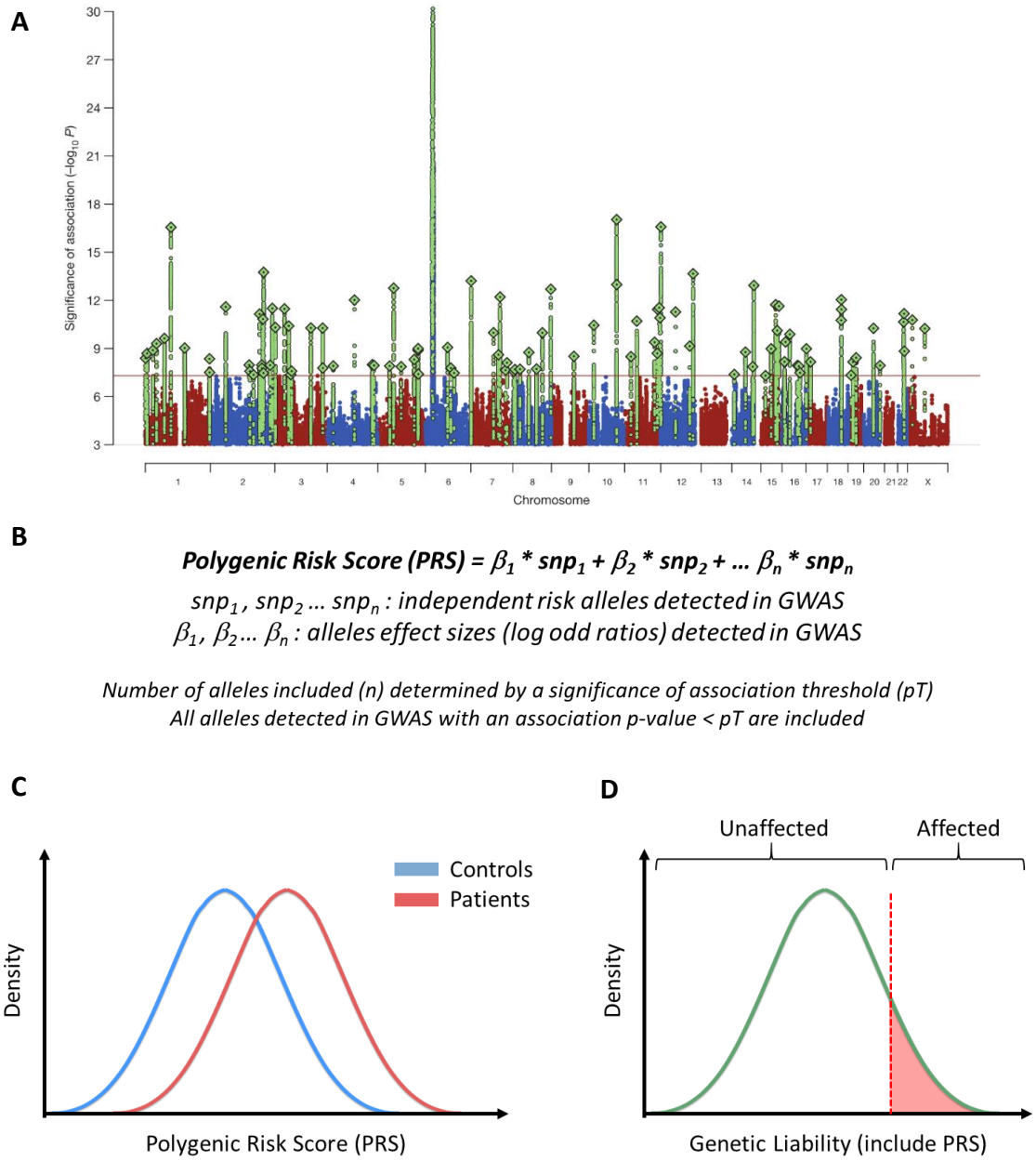


Figure 2-1. Polygenic risk score: theory overview. (A) Genome-Wide Association Study (GWAS) results: Manhattan plot showing schizophrenia association. The x axis is chromosomal position and the y axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level ($5*10^{-8}$). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations. Figure reprinted (adapted legend) by permission from Springer Nature: Nature, Ripke *et al.*2014 [92], copyright (2014)(cf. ref for study details). **(B) Polygenic risk score calculation.** The polygenic risk score is the weighted sums of independent risk alleles that have been detected in GWAS such as presented in (A), for a p-value threshold pT. (C) Case/control PRS distribution. The distribution of PRS in a population follow a normal distribution. Cases/patients tested for the disease/disorder-specific PRS will have, on average, a higher PRS even though case and control distributions will overlap. High or low PRS imply high or low probability of developing the disorder. **(D) Genetic liability.** The PRS is part of the total genetic liability that also include other type of variants. The minimum genetic liability sufficient to cause the disorder (liability threshold) is represented by a red dotted bar.

the genetic risk associated with a particular SNP allele. The risk for all variants with an

association p-value below a pre-defined significance level is considered in aggregate to constitute the PRS, as shown Figure 2-1.B. This allows a PRS for a given disease to be calculated for each individual. For a polygenic disorder, the average PRS will be expected to be higher for the case population (with the disorder) than for the controls (Figure 2-1.C). PRS is typically calculated using SNPs selected at a range of p-value thresholds, and while lower thresholds will include less false-positive SNPs, their smaller numbers will limit the power of detecting differences in the PRS between cases and controls.

2.1.3. Polygenic risk score and schizophrenia

Polygenic risk score has been successfully applied to schizophrenia GWAS data [202] when it was demonstrated that, despite failing to meet the stringent genome-wide association criteria, a large proportion of genetic loci accounted for a consequential proportion of variation in risk between individuals. A more recent analysis of common variants in a larger, more powerful GWAS determined that around 7% of variance in liability can be explained by PRS, with around half of it due to genome-wide significant loci [92].

2.1.3.1. Schizophrenia PRS and schizophrenia aetiology

The schizophrenia PRS has been used to get a better understanding of the aetiology of this disorder. In children (7-9 years), this risk score is significantly associated with a range of cognitive, social, behavioural and emotional deficits that are linked with schizophrenia [203,204]. Schizophrenia PRS has also shown association with negative syndromes and anxiety disorders during adolescence (12-18 years), but not positive symptoms such as psychotic experiences [205]. Moreover, children with child-onset schizophrenia, a severe form of the disorder with an age of onset before 13 years of age, have a higher polygenic risk score than their healthy siblings [206]. This suggests the importance of the schizophrenia PRS to understand the disorder and possibly try and predict its onset. In individuals with first-episode psychosis, PRS has been shown to significantly discriminate individuals that were later on diagnosed with either schizophrenia or other psychotic disorders [207]. The accuracy of this method is however still limited. The PRS of individuals with schizophrenia that reported a family history of psychotic illness have been reported to be higher than the ones without such history [208].

This could indicate a more familial subtype of schizophrenia, in which case in the future PRS screening could potentially help early diagnosis of schizophrenia.

2.1.3.2. Schizophrenia PRS and brain imaging studies

The joint analysis of brain imaging techniques and polygenic risk score is also helping to understand schizophrenia and its endophenotypes. Higher PRS for schizophrenia have been significantly associated with reduced total brain [209] and white matter volumes [209,210], as well as greater globus pallidus volumes [211]. However the association with white matter volume was not replicated in another study [212]. Moreover, larger studies testing association between PRS and brain volumes in healthy individuals did not show significant results [213,214]. In patients with first-episode psychosis or at-risk mental state individuals, higher PRS is associated with reduced hippocampal volume, a known marker of these disorders [215]. Functional imaging also revealed that increased PRS was associated with multiple deficits such as altered brain activation in the ventral striatum and the frontal pole during probabilistic decision-making (choice behaviour) in healthy individuals [216] and neural inefficiency in the left dorsolateral prefrontal cortex in schizophrenia patients [217].

2.1.3.3. Schizophrenia PRS and cross-disorder associations

Some of the common variants that have been associated with schizophrenia have also been shown to be associated with other disorders. For instance, schizophrenia PRS is able to significantly discriminate individuals with bipolar disorder from controls, indicating that both psychiatric disorders share a genetic component [202]. These results have since been replicated [218]. Further studies showed that some genetic loci can be associated with multiple psychiatric disorders. A cross-study of five major psychiatric disorder (Attention Deficit-Hyperactivity Fisorder (ADHD); Autism Spectrum Disorders (ASD); Bipolar Disorder (BPD); Major Depressive Disorder (MDD); Schizophrenia) showed that the schizophrenia PRS discriminate controls from patients with either BPD, MDD or ASD (but not ADHD) [219]. Conversely, PRS analyses based on either BPD, MDD or ASD GWAS summary statistics were able to discriminate between controls and schizophrenia patients. Moreover, the BPD PRS is associated with clinical symptoms in schizophrenia cases (namelymanic symptoms) [220]. It suggests that the severity of the disorder could be influenced by common risk variants. Another study also presented a significant

association of schizophrenia PRS with obsessive-compulsive disorder (OCD) status [221]. These cross-disorder associations with polygenic risk scores suggest that common biological pathways are altered within these different disorders, such as voltage-gated calcium-channel signalling [219].

2.1.3.4. Schizophrenia PRS and 22q11.2 deletion syndrome

A study by Merico *et al.* in 2015 looked at a possible association of schizophrenia PRS and schizophrenia status in 22q11.2 deletion syndrome patients [160]. No significant association were found (despite a trend toward significance) but the number of samples tested were quite limited (6 individuals with schizophrenia versus 3 without). Tansey *et al.* have tested if the schizophrenia PRS was associated with schizophrenia status in individuals with a schizophrenia-associated copy-number variant (CNV) and found a significantly higher PRS compared to controls [161]. Similarly, this study was underpowered to test association for individual CNVs. The results were however consistent with a liability threshold model (Figure 2-1.D) in which the CNVs add to the cumulative risk alongside other alleles such as common variants. Indeed, they reported that schizophrenia cases carrying a high schizophrenia odd-ratio CNV (such as 22q11.2) have a lower PRS than schizophrenia cases carrying a low odd-ratio CNV or no schizophrenia-associated CNV. In this model, the increased risk due to the presence of the CNV means that a lower PRS is needed to pass the liability threshold.

2.2. Aims of the chapter

The work described in this chapter set out to determine if 22q11.2 deletion carriers with schizophrenia have a significantly greater schizophrenia PRS than deletion carriers with no psychotic symptoms. To achieve this, schizophrenia PRS was calculated in the largest to date (to our knowledge) sample of 22q11.2 deletion syndrome patients. This will determine if the aetiology of schizophrenia in 22q11.2 deletion syndrome is due, at least in part, to common risk variants associated with the disorder. To follow up this analysis, the odds ratio conferred by schizophrenia PRS in 22q11.2 deletion carriers was compared to that seen in idiopathic schizophrenics and population controls who did not carry a CNV at 22q11.2. The aim of this secondary analysis was to help investigate if schizophrenia in 22q11.2 deletion syndrome followed a liability threshold model.

2.3. Methods

2.3.1. Cohorts

2.3.1.1. CLOZUK cohort

In the following study, the "CLOZUK" samples refer to the cohort of idiopathic schizophrenia cases and general population controls. The UK Multicentre Research Ethics Committee approved the study and all samples were from participants who provided written informed consent (unless specified otherwise). Access to the dataset was provided by Mark Einon (Cardiff University) with the authorisation of Dr. James Walters (Cardiff University). It is defined as the following cases and controls:

- Schizophrenia cases (named idiopathic schizophrenia samples) include samples from the CLOZUK1, CLOZUK2, COGS1 and COGS2 schizophrenia studies and have been described before [92,222–224]. CLOZUK1/2 cases were diagnosed with treatment-resistant schizophrenia (TRS) and were taking Clozapine, an antipsychotic drug used to treat TRS and owned by Novartis, who provided CLOZUK1 blood samples. CLOZUK2 cases blood samples were collected in collaboration with Leyden Delta (a company involved in the supply and monitoring of Clozapine in the UK) as part of the European Union Seventh Framework Programme (EU-FP7) study CRESTAR. CLOZUK1/2 samples were collected anonymously in the across the United Kingdom without express patient consent consistently with the UK Human Tissue Act. The COGS cohorts (Cardiff Cognition in Schizophrenia study) include schizophrenia patients recruited in Wales and England from voluntary, community and inpatient sector mental health services. CLOZUK1 cases were genotyped at the Broad Institute (Massachusetts, USA) on Illumina HumanOmniExpress-12v1 and HumanOmniExpressExome 8v1 arrays; CLOZUK2 and COGS cases were genotyped by deCODE Genetics (Reykjavík, Iceland) on Illumina HumanOmniExpress-12v1-1_B arrays. These samples have been used in other studies [92,139,161,222,223,225–227] and the validity of use of TRS cases as schizophrenia samples have been demonstrated [225,227].

- Controls are from publicly available datasets (dbGaP) and from the Wellcome Trust Case Control Consortium 2 (www.wtccc.org.uk/info/access_to_data_samples.html). An additional 900 controls (unscreened for psychiatric illness) were recruited from the UK National Blood Transfusion Service by Cardiff University. All controls were genotyped on Illumina Omni arrays.

2.3.1.2. 22q11.2 IBBC cohort

The 22q11.2 IBBC (International 22q11.2 Deletion Syndrome Brain Behaviour Consortium) cohort has been described in Gur *et al.* 2017 [228] and Bassett *et al.* 2017 [158]. Subjects were recruited from 22 international sites from Canada, USA and Europe and have provided informed consent. The study was approved by the local institutional research ethics board of each site. DNA samples were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0 at the Albert Einstein College of Medicine (New York, USA). The presence and extent of the 22q11.2 deletion in individuals has been assessed by the consortium prior to analysis.

In the present study only individuals older than 25 years, with or without schizophrenia, were kept for analysis (excluding individuals with non-schizophrenic psychotic disorders). Details of case/control sample sizes (after quality control) and their origin (international sites participating in the 22q11.2 IBBC) are described in Supplementary table 2-1.

2.3.2. Data processing / Quality control (QC)

The details of the data processing and quality control (QC) are described in Supplementary table 2-2. QC was performed with PLINK 1.9 [229,230] (https://cog-genomics.org/plink2) and in R (Microsoft R Open 3.4.0 (Microsoft), based on R-3.4.0 (R Statistics)). Note: 22q11.2 IBBC samples not analysed in this study (non-schizophrenic psychotic disorders patients and individuals < 25 y.o.) were processed at the same time to use for other analyses not presented here.

A first serie of steps was performed on 22q11.2 IBBC data to prepare for imputation. Genotypes were called from raw intensity data with the Birdseed v2 genotyping algorithm [231] integrated within the Affymetrix Genotyping Console (http://affymetrix.com). After confirming that the genotyped gender was concordant with the gender information available, non-autosomal SNPs, SNPs with no name or with a null allele were removed. Then low-quality SNPs (present in less than 95% of samples) and low-quality samples (with less than 97% of SNPs correctly genotyped) were removed. Genome-wide Identity-By-Descent (IBD) was estimated to remove duplicate samples and determine cryptic relatedness (3rd degree relative or closer removed, PI_HAT > 0.125); for each pair of related samples the exclusion was done based on phenotype (cases prioritised over

controls) and then sample missingness (sample with lowest missingness kept). Population stratification analysis was performed by principal component analysis (PCA) after merging the data with 1000 Genomes Project data Phase 3 (www.internationalgenome.org/data/; named 1000 Genomes data below) that contains genotyped samples from different ethnicities (PCA plots in Supplementary figure 2-1.A, 1000 Genomes data population details in Supplementary table 2-3; only the Finish, Han Chinese (Beinjing) and Yoruba population are used for this first PCA). All PCAs were done using the *smartpca* function of EIGENSTRAT (EIGENSOFT package) [232]. Samples that were further than 2 standard deviations from the mean of any principal components were excluded. Sample heterozygosity was checked, then SNPs with a Hardy-Weinberg equilibrium value HWE < $10e^{-5}$ have been excluded (with "mid-p" adjustment). The resulting dataset have been then imputed on the Michigan Imputation Server by Dr. Leon Hubbard, Cardiff University (Eagle V2.3 [233] was used for pre-phasing, with the HRCv1.1 as the imputation reference panel [234]).

The CLOZUK dataset has been provided already imputed, as described in Pardinas *et al.* [225]. The CLOZUK dataset was phased and imputed using the SHAPEIT/IMPUTE2 [235,236] algorithms with a combination of the 1000 Genomes phase 3 (1KGPp3) and UK10K datasets as reference panel, while the 22q11.2 dataset has been phased and imputed using the Eagle2/Minimac3 algorithms [237] with the larger HRCv1.1 reference panel as it has been shown to allow for greater imputation accuracy than the two other reference panels [234]. However, for common variants, both imputation algorithms have been shown to have similar accuracy regardless of the reference panel used [234,237]. While the use of different imputation methods might lead to spurious association in variant discovery, it is less likely to affect the results of polygenic risk scoring which takes into account SNPs with small effect sizes all across the genome as a sum. Moreover, stringent quality control has been performed on both datasets before merging to obtain good quality data [238].

After imputation, only SNPs with high confidence (imputation information score > 0.9) were kept, and then filtered by missingness (< 1%) and HWE (< 10e-5, with "mid-p" adjustment) for both 22q11.2 and CLOZUK datasets. The two datasets were then merged based on overlapping SNPs. SNPs present in less than 99% of samples in the merged dataset were removed, and an IBD analysis was performed to remove 3rd degree relatives and potential individuals present in both 22q11.2 and CLOZUK dataset. A

	22q11 SCZ	22q NonAffect	Idiopathic SCZ	Controls
Sample size	105	171	10755	24295
Gender ratio M/F	0.91	0.55	2.47	1.04

Figure 2-2. Sample size and gender ratios for all groups used in analysis. 22q SCZ: 22q11.2 deletion syndrome patients with schizophrenia (≥ 25 y.o.); 22q NonAffect: 22q11.2 deletion syndrome individuals without psychotic disorders (≥ 25 y.o.); Controls: healthy control population individuals; Idiopathic SCZ: individuals with schizophrenia. Unless specified otherwise all the samples shown here are included in the analyses.

second population stratification analysis was done and samples further than 6 standard deviations away from the mean were removed (PCA plots in Supplementary figure 2-1.B, 1000 Genomes data population details in Supplementary table 2-3). Samples from the CLOZUK cohort with a known pathogenic copy number variant (associated with schizophrenia, as characterised by Dr. Elliott Rees, Cardiff University [223]) were removed, as well as SNPs included within the genomic region 22q11.2. Finally, SNPs with a minor allele frequency < 10% were removed prior to polygenic profiling. This final dataset is called in the following section the target sample. The number of individuals retained after QC are detailed Figure 2-2.

2.3.3. Polygenic risk score calculation

The polygenic risk score (PRS) of each individual has been calculated with the PRSice software [239]. The analysis pipeline is described in Figure 2-3. We used as a discovery sample (training dataset) the summary statistics of a version of the latest Psychiatric Genomics Consortium (PGC) schizophrenia meta-analysis [92] that does not include samples used in the target sample (CLOZUK idiopathic schizophrenia samples and controls). These results generated by the PGC excluding CLOZUK samples are part of a leave-one out GWAS provided to the members of the consortium (Professors MJ Owen and MC O'Donovan), enabling independent discovery and replication data-sets for downstream analysis. While the 22q11.2 IBBC samples have not been molecularly compared for overlapping samples yet, the IBBC groups have all declared that they have not contributed any of these samples to the PGC. The discovery sample SNPs were selected at different significance of association p-value thresholds ($pT < 0.0001$, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1) and then clumped to include SNPs in linkage disequilibrium of $R^2 > 0.1$ within 250kb windows. The polygenic risk scores were created

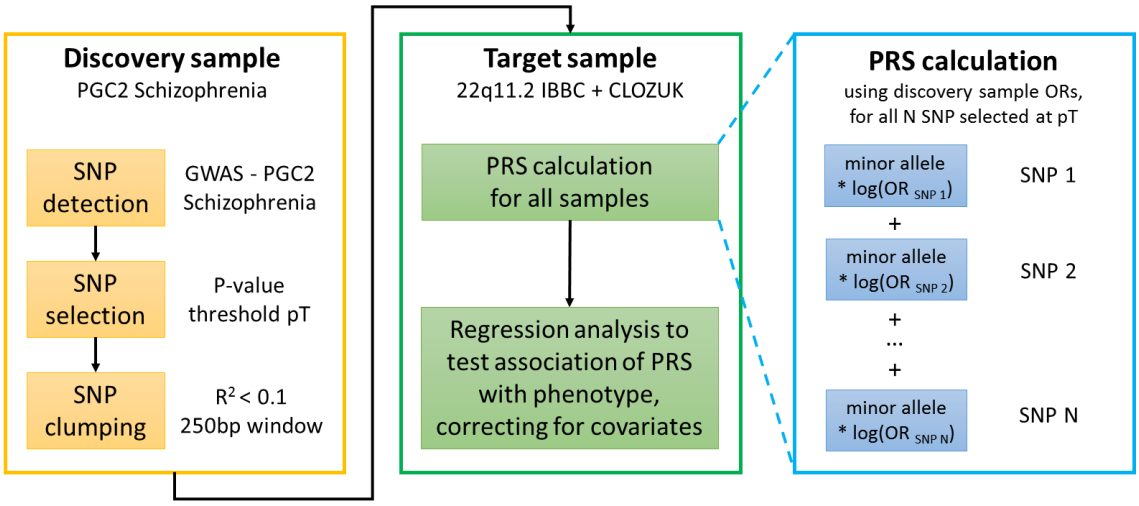


Figure 2-3. Overview of the analysis pipeline. The SNP selection, SNP clumping and PRS calculation steps were performed with PRSice software (REF). OR: Odd ratio; PRS: Polygenic Risk Score; R^2: Linkage Disequilibrium R^2 value; SNP: Single Nucleotide Polymorphism. Adapted from: SlideShare: "Analysis of the Genetic Overlap of Borderline Personality Disorder and Bipolar Disorder" by: Prof. Stephanie Witt (Central Institute of Mental Health, Mannheim) (*https://www.slideshare.net/ISBD/analysis-of-the-genetic-overlap-of-borderline-personality-disorder-and-bipolar-disorder*).

at each of the p-value thresholds pT for each sample by summing the GWAS log odd ratios of the included SNPs. SNPs within the MHC region were excluded from analysis due to the long range linkage disequilibrium in this region [240].

2.3.4. Statistical analysis

All statistical analysis was performed in R (Microsoft R Open 3.4.0 (Microsoft), based on R-3.4.0 (R Statistics)). The custom scripts used will be made available on GitHub (github.com/MonfeugaT) after publication of the results. Logistic regressions were used to determine the association between the PRS and phenotype status. A principal component analysis was run on the target sample and the first 20 principal components, as well as gender, were used as covariates in all statistical models. The 20 principal components used as covariates are represented in Supplementary figure 2-2. 4 different groups were defined and compared in the different analyses: 22q NonAffect: 22q11.2 deletion syndrome patients without schizophrenia; 22q SCZ: 22q11.2 deletion syndrome patients with schizophrenia; Controls: control population individuals without known psychosis; Idiopathic schizophrenia: individuals with schizophrenia but without 22q11.2 deletion.

The proportion of variance in the model explained by polygenic risk score, defined by the Nagelkerke R^2, was obtained by subtracting the R^2 of a full model including PRS and covariates (Phenotype ~ PRS + covariates, with Phenotype as a binary value designing the two group from a pairwise comparison) from the R^2 of a base model only including covariates (Phenotype ~ covariates). Covariates include the first 20 principal components to account for population stratification and gender. P-values represent the association of the PRS with phenotype for each p-value threshold pT tested.

Schizophrenia odds ratios and 95% confidence intervals have been calculated to represent the effect of PRS on schizophrenia risk, between 22q11.2 samples with or without schizophrenia as well as between idiopathic schizophrenia and controls. They were calculated using an adapted version of the corresponding function from the PRSice software R script, after dividing the samples into different PRS quantiles. Odds ratio and confidence intervals have been calculated with a logistic regression of phenotype on PRS decile that include gender and 20 principal components as covariates to correct for population stratification. Statistical comparison of odds ratio between studies, for each quantile tested, was performed with the *rma()* function of the *metaphor* package in R [241] (comparison of the log odd ratios).

No correction for multiple comparisons has been done to account for the multiple p-value thresholds pT used due to the high correlation of results between thresholds. No correction was applied either to correct for the multiple group pairwise comparisons. However, a permutation resampling analysis was performed to control for Type I error. For each pairwise comparison, the phenotypes were randomised while keeping the same case/control ratio. The p-values shown Supplementary figure 2-4 represent the probability of obtaining a smaller p-value by chance after random sampling (N = 10,000 resampling). These p-values are described below as P_{perm}, while the p-values represented in Figure 2-4.A and described as P below are obtained from the logistic regression analysis.

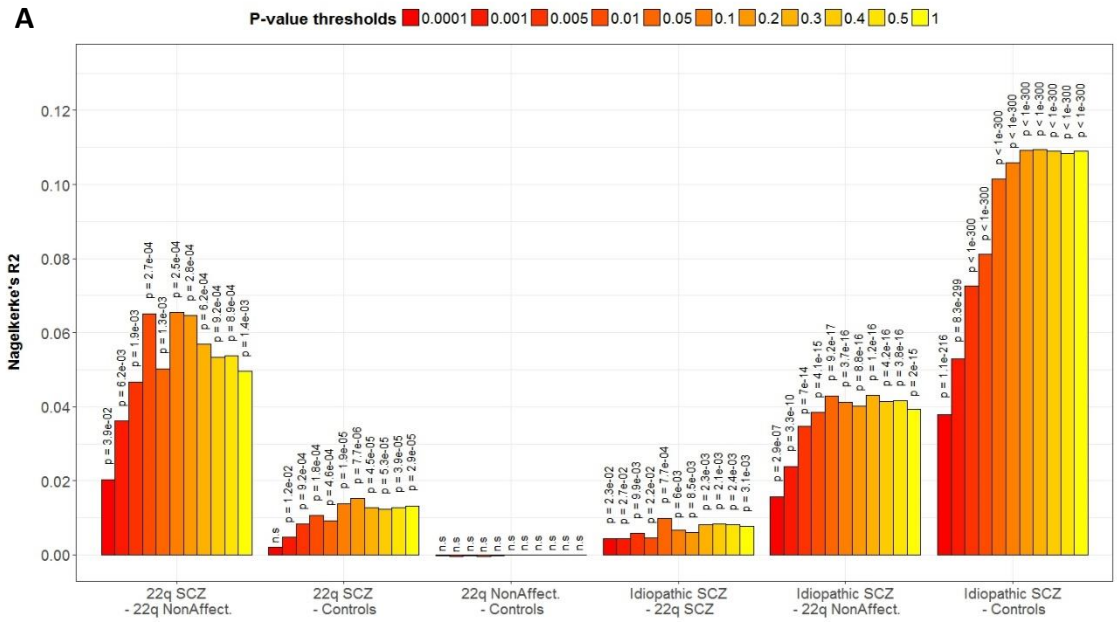
Plots were generated with the *ggplot2* package [242].

2.4. Results

2.4.1. Proportion of variance explained by polygenic risk score

Following quality control analysis, the study was left with 105 schizophrenics who carried a 22q11.2 deletion (22q11.2 IBBC cases), 171 deletion carriers without

A



B

pT	0.0001	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4	0.5	1
# SNPs	342	952	2226	3244	8379	12712	19360	24684	29138	32941	45713

Figure 2-4. Proportion of variance in schizophrenia explained by polygenic risk score (PRS). (A) Plots of Nagelkerke's R^2 for each pairwise comparison and different p-value thresholds. Nagelkerke's R^2 represent the proportion of variance in schizophrenia explained by the polygenic risk score and is calculated by substracting the R^2 of a regression model including PRS and covariates from the R^2 of a model including covariates only (gender and 20 principal components to control for population stratification). The p-value represented on top of bars represent the association of PRS with the phenotype in the model for each p-value threshold (shown by colors, cf. legend on top). For each "Group A – Group B" comparison, a positive R2 indicate that Group A is associated with an increased risk for schizophrenia. n.s.: non-significant. **(B) Number of independent SNPs included in the analysis for each p-value threshold pT**.

schizophrenia (22q11.2 IBBC controls), 10,755 samples with idiopathic schizophrenia (CLOZUK cases) and 24,295 population controls (CLOZUK population controls) (Figure 2-2). Pairwise-comparison of polygenic risk scores was performed between the samples of the 4 groups. The proportion of variance in schizophrenia explained by the PRS in the model between two groups is defined by the Nagelkerke's R^2, as represented Figure 2-4.A.

2.4.1.1. Comparison of schizophrenia PRS between idiopathic CLOZUK cases and CLOZUK population controls

First, individuals with idiopathic schizophrenia and no 22q11.2 deletion (CLOZUK cases) had a significantly higher polygenic risk score than the CLOZUK population controls. This was obtained at all pT thresholds and is consistent with what has been reported previously [92] (Figure 4). At the p-value threshold pT < 0.1 (used for the following descriptions as the representative threshold), the PRS explained 11% of the variance between idiopathic schizophrenia and controls samples (pT < 0.1: $R^2 = 0.11$, P = 0, $P_{perm} < 1*10^{-4}$) (Note: p = 0 means that the P-value is too small to be computed by the R script used; $p < 10*10^{-300}$).

2.4.1.2. Comparison of schizophrenia PRS between 22q11.2 IBBC cases, 22q11.2 IBBC controls and CLOZUK population controls

The 22q11.2 deletion carriers with schizophrenia (22q11.2 IBBC cases) had a significantly higher schizophrenia PRS than the 22q11.2 deletion carriers who did not have schizophrenia (22q11.2 IBBC controls) (pT < 0.1: $R^2 = 0.065$, $p = 2.5*10^{-4}$, $P_{perm} = 2*10^{-4}$). Similarly, the 22q11.2 IBBC cases also had a significantly higher PRS than the CLOZUK population controls who do not carry the 22q11.2 deletion (pT < 0.1: $R^2 = 0.014$, $p = 1.9*10^{-5}$, $P_{perm} < 1*10^{-4}$). Schizophrenia PRS was therefore able to significantly distinguish the 22q11.2 carriers with schizophrenia (22q11.2 IBBC cases) and both non-schizophrenic groups (22q11.2 IBBC controls and CLOZUK population controls). No differences were observed between the two groups of non-psychotic individuals (pT < 0.1: $R^2 = -5.7*10^{-5}$, p = 0.74, $P_{perm} = 0.73$).

2.4.1.3. Comparison of schizophrenia PRS between CLOZUK cases and 22q11.2 IBBC cases and 22q11.2 IBBC controls

The CLOZUK cases had a greater PRS than both groups that carried a 22q11.2 deletion (IBBC cases and controls). The difference was highly significant when comparing to the IBBC controls (pT < 0.1: $R^2 = 0.041$, $p = 3.7*10^{-16}$, $P_{perm} < 1*10^{-4}$), but much less when comparing to IBBC cases (pT < 0.1: $R^2 = 0.0066$, $p = 6.0*10^{-3}$, $P_{perm} = 5.4*10^{-3}$).

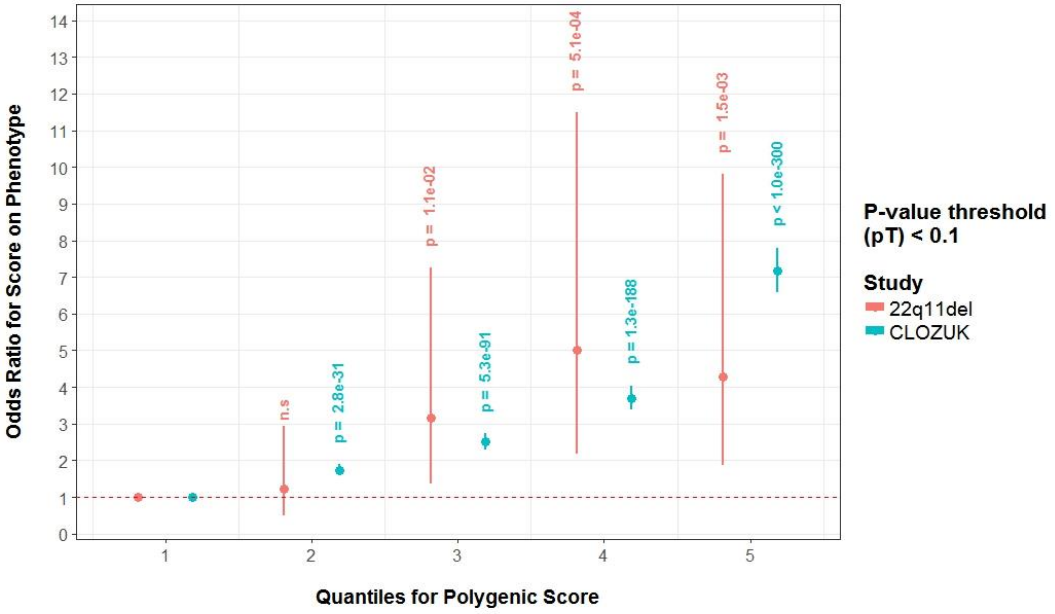


Figure 2-5. Odds ratio by polygenic risk score. Plots of odds ratio for schizophrenia by polygenic risk score (PRS) quantile with or without the presence of 22q11.2 deletion for p-value threshold pT < 0.1. Odds ratio are presented by PRS quintiles. Odds ratio for schizophrenia calculated for samples that have the 22q11.2 deletion (with or without schizophrenia) are represented in red while those calculated for samples without it (CLOZUK/WTCCC cohorts) are represented in blue. Bars represent the 95% confidence interval. Odds ratio and confidence intervals have been calculated with a logistic regression of phenotype on PRS decile that include gender and 20 principal components as covariates to correct for population stratification. Significance for difference from the 1st decile is represented above confidence interval bars (for each study separately). The number of independent SNPs included at each p-value threshold is detailed in Figure 2-4.B.

2.4.2. Estimating the odds ratio for schizophrenia conferred by the polygenic risk score

22q11.2 deletion carriers (IBBC cases and controls), idiopathic schizophrenics and population controls (CLOZUK cases and controls) were grouped and ranked into polygenic risk score quintiles and, with reference to the lowest quintile, the odds ratio for schizophrenia was calculated for each quintile. Only the p-value threshold pT < 0.1 has been tested, as it's the threshold for which IBBC cases and IBBC controls are the most significantly discriminated (Cf. Figure 2-4). As it has been published previously [92], the odds ratios seen in idiopathic schizophrenics were greater in samples that had a higher PRS (*i.e.* those who carry a greater number of common variants associated with an increased risk for schizophrenia) (Figure 2-5). There was a similar trend in the IBBC cases; however, compared to the CLOZUK cases, the odds ratio means were greater for individuals with PRS in intermediates quantiles (3rd and 4th quintiles), *i.e.* lower PRS

result in larger odds ratios for schizophrenia in 22q11.2 deletion carriers compared to idiopathic schizophrenics (Figure 2-5). For example, at the p-value threshold pT < 0.1, the odd ratios for the fourth PRS quintile (Figure 2-5) is 5.01 (95% CI [2.18-11.48]) for the IBBC samples and 3.69 (95% CI [3.39-4.02]) for the CLOZUK samples. However, confidence intervals are overlapping between odd ratios for CLOZUK and IBBC samples and their difference is not significant (p=1 for each quintile).

2.5. Discussion

This study presents a polygenic risk score analysis for schizophrenia risk in the largest cohort of 22q11.2 deletion syndrome patients available to date. The goal was to determine if common variants previously associated with schizophrenia in genome-wide associations studies were involved in the aetiology of this psychiatric disorder in 22q11.2DS. Results show that, indeed, 22q11.2 deletion carriers with schizophrenia have on average a greater polygenic risk score than 22q11.2 deletion carriers without psychosis ($pT < 0.1$: $R^2 = 0.065$, $p = 2.5*10^{-4}$, $P_{perm} = 2*10^{-4}$). The polygenic risk score explains around 6% of the variance in schizophrenia in 22q11.2DS patients in the model studied, with a mean Nagelkerke's R^2 of 0.051 between all p-value thresholds tested.

The CLOZUK1/2 dataset is to our knowledge currently the largest available dataset of idiopathic schizophrenia patients and controls independent from the PGC meta-analysis (CLOZUK1 were included in the original PGC meta-analysis [92] but new summary statistics have been generated without them while CLOZUK2 samples were not in the original dataset). This cohort was used to assess the relative contribution of schizophrenia PRS in 22q11.2 deletion carriers to that seen in idiopathic schizophrenics and unscreened population controls. The results show that the 22q11.2 deletion carriers with schizophrenia had a PRS significantly higher than the control population ($pT < 0.1$: $R^2 = 0.014$, $p = 1.9*10^{-5}$, $P_{perm} < 1*10^{-4}$) and significantly lower than the idiopathic schizophrenia patients ($pT < 0.1$: $R^2 = 0.0066$, $p = 6.0*10^{-3}$, $P_{perm} = 5.4*10^{-3}$). This observation is consistent with an additive model where both the 22q11.2 deletion and schizophrenia polygenic risk contribute to the overall genetic liability of the increased rate of schizophrenia in 22q11.2DS.

In such a model, individuals carrying the deletion, a known risk factor for schizophrenia [223], require a lower polygenic risk score (relative to idiopathic

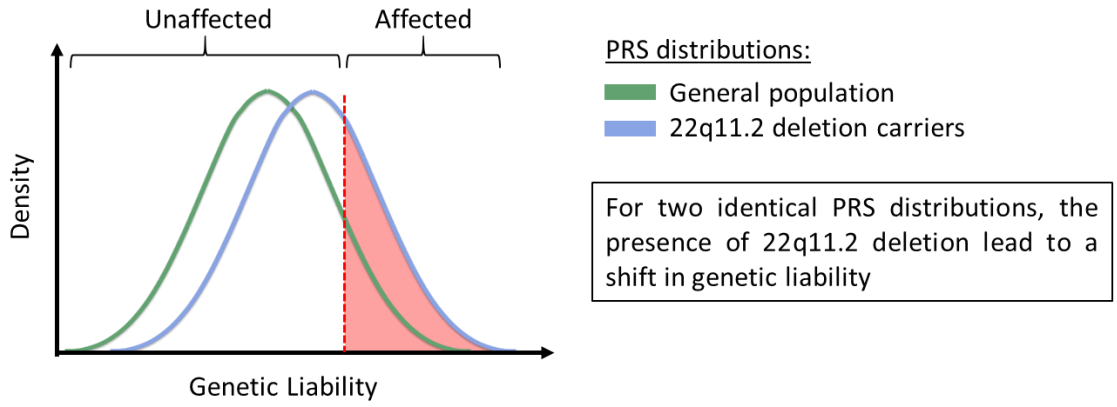


Figure 2-6. Effect of 22q11.2 deletion on genetic liability. The 22q11.2 deletion participate to the genetic liability and will lead to a shift of the population curve. In consequence, polygenic risk score will pass the threshold with lower values than in a population with individuals that do not have the deletion. The minimum genetic liability sufficient to cause the disorder (liability threshold) is represented by a red dotted bar.

schizophrenics) to exceed the liability threshold required to develop schizophrenia (Figure 2-6). This is further supported by a comparison of the odds ratios for schizophrenia conferred by the PRS in 22q11.2 deletion carriers (IBBC cases) and idiopathic schizophrenics (CLOZUK cases). As presented in Figure 2-5 the odds ratios for intermediate PRS quantiles are greater for 22q11.2 deletion carriers relative to idiopathic schizophrenics, implying that relatively lower polygenic risk scores can confer a greater risk to schizophrenia in the presence of the 22q11.2 deletion. However, the odds ratios confidence intervals are overlapping, and the difference isn't significant; greater sample size for the 22q11.2DS population would reduce the confidence intervals and help further comparisons of both groups.

As high-to-intermediate PRS values could therefore be sufficient to cause schizophrenia in 22q11.2 deletion carriers, conversely it might also be expected that 22q11.2 deletion carriers who do not develop schizophrenia to have on average a lower PRS than that seen in population controls. However, this study was not able to demonstrate a significant difference between the IBBC controls and the CLOZUK population controls ($pT < 0.1$: $R^2 = -5.7*10^{-5}$, $p = 0.74$, $P_{perm} = 0.73$). It is possible that this could be due to insufficient power, and if so then an increased sample size will be needed in the future to determine if it is possible to demonstrate a significant difference in schizophrenia PRS between non-schizophrenic individuals carrying a 22q11.2 deletion and population controls.

Schizophrenia PRS explained a higher percentage of the variance in this model when comparing the IBBC cases and IBBC controls ($pT < 0.1$: $R^2 = 0.065$, $p = 2.5*10^{-4}$, $P_{perm} = 2*10^{-4}$) than when the IBBC cases and CLOZUK population controls were compared ($pT < 0.1$: $R^2 = 0.014$, $p = 1.9*10^{-5}$, $P_{perm} < 1*10^{-4}$) (Figure 2-4, Supplementary figure 2-4). The difference in variance in schizophrenia explained by PRS in this model could be due to the deletion, as in one comparison (IBBC cases and CLOZUK population controls) some of the variance in schizophrenia will be explained by the effect of the deletion itself. This difference could be due to a direct effect of the deletion but also to an interaction effect between the deletion and common variants.

The PRS explained a lower percentage of the variance in this model when comparing the IBBC cases and IBBC controls ($pT < 0.1$: $R^2 = 0.065$, $p = 2.5*10^{-4}$, $P_{perm} = 2*10^{-4}$) than that seen when the CLOZUK cases and CLOZUK population controls ($pT < 0.1$: $R^2 = 0.11$, $P = 0$, $P_{perm} < 1*10^{-4}$) were compared (Figure 2-4, Supplementary figure 2-4). As one test compares both groups with a 22q11.2 deletion and the other includes no deletion carriers, the difference in variance in schizophrenia might be due to an interaction effect of the 22q11.2 deletion with common variants, rather than a direct effect. Indeed, in the comparison of IBBC cases and IBBC controls the deletion is not part of the variance in schizophrenia as it is present in both cases and controls. In consequence, it is possible that an epistatic effect between PRS and deletion is taking place (interaction between the deletion and common variants).

Careful interpretation of these results is needed; however, they support an additive model where both the deletion and common variants are part of the genetic liability. This observation is concordant with an earlier study of PRS in schizophrenia-related CNVs (as a group) [161].

There are differences in gender ratios (male/female) between the 4 groups tested, as shown Figure 2-2. However, performing permutation analysis while keeping similar gender ratio between groups shows that it does not seem to affect the results (Supplementary figure 2-3). Furthermore, despite unequal sample sizes between the groups, the R^2 values and their differences between pairwise comparison is consistent when performing random sampling of equal sample size in each group, as shown in Supplementary figure 2-5.

The number of SNPs used to detect PRS differences between the IBBC cases and controls was relatively low (Figure 2-4.B) due to the joint data processing and quality control with the CLOZUK data cohort, which was genotyped using different array technology. While genotype imputation helped to improve the number of overlapping variants between the different types of arrays used in the two cohorts (CLOZUK/22q11.2 IBBC), the number of high quality imputed SNPs remained low after QC. Processing the 22q11.2 IBBC data separately yielded around twice as many SNPs tested at each p-value threshold (Supplementary figure 2-6.C) and resulted in similar levels of variance explained by the PRS (Supplementary figure 2-6.A). Another possible concern would be the multiple origins of the samples collected within the 22q11.2 IBBC cohort, that could lead to an ascertainment bias. For this reason, the sample origin has been added to the regression model to compare 22q11.2DS samples with or without schizophrenia (for the 22q11.2 dataset processed independently). No major differences were observed when correcting for this potential bias (Supplementary figure 2-6.B).

Increased sample sizes and further investigation will be needed to characterise the difference in variance in schizophrenia explained by PRS with or without the presence of the deletion. The polygenic risk score tends to underestimate the total variation in liability explained by the PRS because of errors in the estimation of variants effects [243,244]. For this reason, it is different than the Nagelkerke's R^2 calculated in the present study. Future work will focus on trying to estimate this variance explained by all schizophrenia-associated SNPs in 22q11.2 by using software such as GCTA [245]. Another future goal will be to try and explore possible epistatic interactions between the deletion and other variants. Some genes located within the 22q11.2 deletion could play a role in such interactions, such as DGCR8, a protein involved in genetic regulation (cf. Chapter 4 (p.81)).

3. Effect of the 22q11.2 deletion on DNA methylation

3.1. Introduction

3.1.1. DNA methylation and gene expression

DNA methylation involves a chemical modification of DNA bases and is one of the most studied epigenetic modifications in animals, yet it is still not well understood despite having been discovered more than 60 years ago [246]. The most common type of DNA methylation mark is the 5-methylcytosine (5mC) (Figure 3-1.A); it occurs on cytosine bases (C) that are followed by a guanine (G), or CpG sites. In animals, the methylation pattern is bimodal with the vast majority of CpG sites having high levels of methylation (>80%) while particular sites, which are termed "CpG islands", have low levels of methylation (<20%).

As around 70% of gene promoters have CpG islands (CGIs) they have been one of the first and most studied feature of 5mC methylation [247]. Most CGIs are in a nonmethylated state, in particular those located in promoter regions where methylation is often associated with long-term silencing of the gene (as reviewed in Deaton & Bird 2011 [248]). The effect of methylation on expression is context dependent: they are negatively correlated in CpGs close to the transcription start site (TSS) while they are positively correlated in gene body CpGs [249] (Figure 3-1.B).

3.1.2. DNA methylase enzymes

There are three known DNA methylase: DNMT1, DNMT3a and DNMT3b. DNMT1 is commonly considered as a maintenance methylase due to its ability to preferentially process hemimethylated (or asymmetric) CpG sites [250]. This allows the preservation of DNA methylation patterns in a cell lineage despite DNA replication (Figure 3-1.C.1). DNMT3a/b on the other hand are mainly (but not exclusively) involved in *de novo* DNA methylation and can introduce methylation at non-methylated CpG sites [251] (Figure 3-1.C.2). DNMT3l is a fourth member of the DNMT family but is not a methylase *per se* due to its lack of catalytic domain. It is thought to interact with the other DNMT proteins, in particular DNMT3a/b in the establishment of genomic imprints [252], a specific case of DNA methylation leading to a parent-of-origin specific allelic expression. DNA

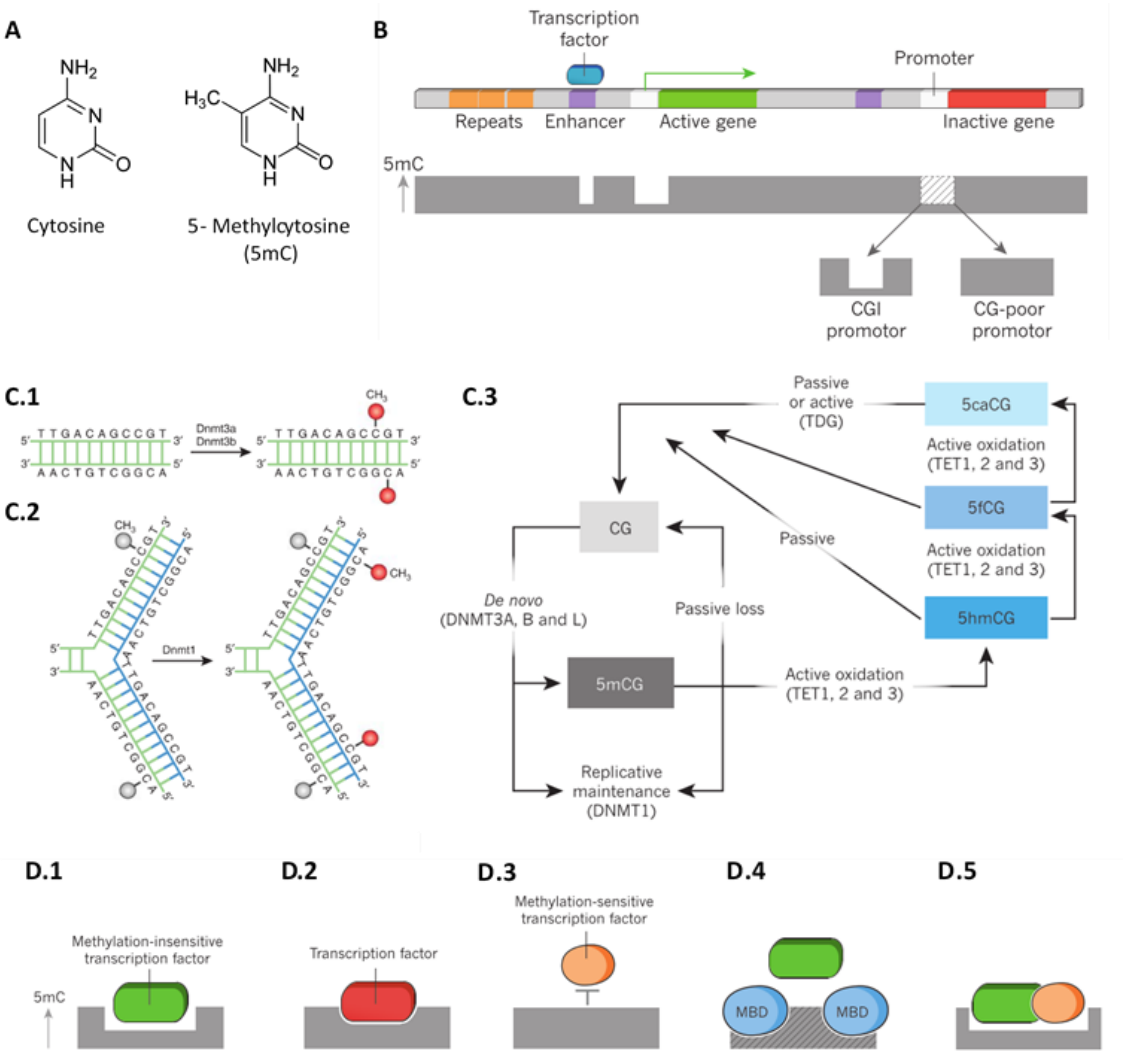


Figure 3-1. Overview of DNA methylation biology. (Legend on next page)

demethylation can occur in either a passive or active manner. DNMTs dysfunction or inhibition can lead to altered DNA methylation patterns in dividing cells due to a lack of maintenance. Even though there is no known DNA demethylase enzyme, active demethylation has been observed via repeated oxidisation and removal of the methyl group (as reviewed in Kohli & Zhang 2013 [253]) (Figure 3-1.C.3). The biological function of DNA methylation and its correlation with gene expression is still unclear and under investigation. In particular, the relation between transcription factor binding and DNA methylation is subject to debate and seems to be context dependent (Figure 1.D). It is however clear that DNA methylation is essential for development; DNMT1, DNMT3a and DNMT3b homozygous knock-outs are lethal in mice (pre-natal death for DNMT1 and DNMT3b, post-natal for DNMT3a) [251,254].

Figure 3-1. Overview of DNA methylation biology. (A) Chemical structures of cytosine and 5-methylcytosine (5mC). (B) Genomic distribution of methylated cytosine in a typical vertebrate genome (shown by level of 5-methylcytosine). The representative genomic region includes an example of an active and an inactive gene with proximal (promoter) and distal (enhancer) regulatory regions. CpG islands, which often overlap with promoter regions, generally remain unmethylated, whereas CG-poor promoters are methylated when not active. **(C) Setting and erasing cytosine methylation.** A family of DNA methyltransferases (DNMTs) catalyses the transfer of a methyl group from S-adenyl methionine (SAM) to the fifth carbon of cytosine residue to form 5-methylcytosine. (**C.1**) DNMT3a and DNMT 3b are the de novo DNMTs and transfer methyl groups (red) onto naked DNA. (**C.2**) DNMT1 is the maintenance DNMT and maintains DNA methylation pattern during replication. When DNA undergoes semiconservative replication, the parental DNA strand retains the original DNA methylation pattern (grey). DNMT1 associates at the replication foci and precisely replicates the original DNA methylation pattern by adding methyl groups (red) onto the newly formed daughter strand (blue). (**C.3**) Different methylation states of the CG dinucleotide and the enzymatic pathways that set, maintain and erase the mark. The pathways leading from the oxidized forms to the unmethylated state are under debate. **(D) Potential scenarios for the interplay between cytosine methylation (shown by level of 5-methylcytosine) and transcription-factor binding**. (**D.1**) A methylation-insensitive transcription factor causes reduced methylation after binding. (**D.2**) A transcription factor binds specifically to the methylated state of its binding site. (**D.3**) A methylation-sensitive transcription factor is blocked by 5-methylcytosine. (**D.4**) Methyl-CpG-binding domain (MBD) proteins bind to the methylated state, leading to indirect repression, which probably requires high local density of CGs (shading). (**D.5**) A methylation-insensitive transcription factor functions as a pioneer factor and creates a site of reduced methylation that allows a methylation-sensitive factor to bind. (C.1/2): reprinted by permission from Springer Nature: Neuropsychopharmacology, Moore & al., 2013 [496], copyright (2012); (B,C.3,D) reprinted by permission from Springer Nature: Nature, Schübeler, 2015 [497], copyright (2015). *5mC(G), 5-methylcytosine(guanine); 5hmC(G), 5-hydroxymethylcytosine(guanine); 5fC(G), 5-formylcytosine(guanine); 5caC(G), 5-carboxylcytosine(guanine); CGI, CpG Island; DNMT, DNA methyltransferase; MBD, Methyl-CpG-binding domain; SAM, S-adenyl methionine; TDG, thymine-DNA glycosylase; TET, ten-eleven translocation family of proteins.*

3.1.3. DNA methylation in neurodevelopment

Spiers *et al.* have observed widespread changes in DNA methylation during the development of the human foetal brain [255]. The assessed CpGs showed a significant reduction of methylation over time, with an enrichment of differentially methylated sites in gene bodies, CGI shores and shelves (regions flanking the CGIs). CGIs seemed to be less dynamic than other genomic regions, which reflects their strong presence over the transcription start site of housekeeping genes [256,257]. Jaffe *et al.* confirmed this global remodelling of the epigenome during brain development, more specifically during the transition from foetal to postnatal life [258]. It might be due to changes in neuronal

composition during this time-period, tying in with the tissue-specific changes observed. Interestingly, both studies show that several schizophrenia-associated genomic loci contain sites that are differentially methylated during development. In fact, around 60% of published GWAS-positive schizophrenia loci are a significant methylation quantitative trait loci (meQTLs) in this developmental data [258]. This supports the hypothesis that schizophrenia is a neurodevelopmental disorder [181].

3.1.4. Genome wide DNA methylation analysis in Schizophrenia

DNA methylation microarrays and Methylated DNA ImmunoPrecipitation sequencing (MeDIP-seq) have allowed the study of genome-wide methylation in large case-control studies. The first genome-wide DNA methylation study of major psychosis, performed in 2008 by Mill *et al.*, revealed that DNA isolated from the post mortem frontal-cortex of schizophrenia patients was differentially methylated relative to controls [259]. Some of the affected genes were involved in biologically relevant pathways, such as neurotransmitter pathways (*e.g. GRIN3B* and *GRIA2* in glutamatergic transmission). More recent studies in larger cohorts have since been performed and report differential methylation in many genes previously associated with schizophrenia as well as potential new candidates [258,260–266].

Due to its better availability than brain tissue, many studies have been using peripheral blood to try and detect methylomic changes due to psychosis. It has been reported that DNMT1 and TET1, both involved in methylation pathways (cf. Figure 3-1.C.3), have increased levels both in brain tissue [267,268] and peripheral blood lymphocytes [268,269] of schizophrenia patients. This suggests a possible alteration of DNA methylation patterns in schizophrenia and that changes observed in the brain could potentially also be observed in other tissues. Methylation changes have indeed been observed in blood from patients with first-episode psychosis [270], including in antipsychotic-naïve patients [271]. This suggest that the DNA methylation changes, whether they are the cause or consequences of changes in gene expression, might have a diagnostic value.

Multiple schizophrenia EWAS from peripheral blood have reported differences in methylation affecting genes involved in pathways relevant to the disorder, such as neuronal development and differentiation [272–274]. Results of longitudinal studies of genomic DNA methylation during psychotic transition such as performed by Kebir et al

Discovery threshold	Type	Study	Tissue	Probes	Genes
p < 1 x 10-7	Discovery	Hannon et al. (Table 1)	Blood	25	16
	Replication	Montano et al. (eTable 4, N = 172)	Blood	0	0
		Aberg et al. (Table 2, N= 65	Blood	NA	0
		Jaffe et al. (Table S10, N = 2104)	Brain	0	1
p < 1 x 10-5	Discovery	Hannon et al. (Table 1)	Blood	1223	949
	Replication	Montano et al. (eTable 4, N = 172)	Blood	2	9
		Aberg et al. (Table 2, N= 65	Blood	NA	3
		Jaffe et al. (Table S10, N = 2104)	Brain	6	61

Table 3-1. Replication of significant case-control differences. "Probes" refer to the exact Illumina 450 k probe by "cg" identifier; "Genes" refer to the nearest gene to each probe or those listed in the tables of the cited papers. Table and legend adapted from Jaffe & Kleinman, 2016 [498], used under CC BY 4.0 license. *References: Hannon et al.*[486]*; Montano et al.*[273] *; Alberg et al.*[499] *;Jaffe et al.*[258].

[274] will also help to understand what are the changes occurring during the conversion from the prodromal phase to psychosis.

3.1.5. Considerations when interpreting differential DNA methylation

It has been shown that intra-individual differences between blood and brain tissue are greater than between-individuals differences [275]. Moreover, methylation patterns can be different between brain regions, with sometimes greater differences within subjects than between study groups [264,266]. Consequently, despite there being evidence for a partial overlap (although minor) between the results of schizophrenia EWAS using different tissues [261,276], linking the results of DNA methylation in peripheral tissue to the pathology of psychiatric disease needs to be done with caution. To facilitate this, recent studies have been performed to find which CpG sites methylation levels are correlated between blood and brain tissues [277,278]. This information will provide a great help to interpret peripheral tissue EWAS in the context of psychiatric disease.

Another issue to ponder is the impact of the cell composition of the tissue considered. As the methylation of some CpGs can be specific to their cell-type then when DNAm is measured in DNA extracted from tissue composed of heterogenous cells, it is possible that the variation in cell-type proportions between the samples may result in false observations of differential methylation (or even mask true differential methylation in disease relevant cell types). In blood, the proportions of leukocytes varies with age and immune status; and in the context of schizophrenia it has been shown that antipsychotics such as clozapine induce neutropenia (deficit in neutrophils) [279]. Brain tissue samples are

also heterogeneous and cell composition can vary between patients. Moreover deficits in parvalbumin-positive GABAergic interneurons have been observed in schizophrenia patients [280].

Genome wide analysis of DNAm has resulted in new methods that allow us to use the methylation data to adjust for estimated cellular proportion [281–283]. These have shown that it is essential to use these estimation (or real cell type counts if available) to correct any possibly associated bias [284]. Finally, unlike in certain cancer types were global changes of methylation are observed [285], schizophrenia seems to be associated with more moderate changes for which only larger cohorts will provide enough power to detect properly. Tissue heterogeneity, differences in statistical methods and covariates accounted for as well as lack of power probably explain the low replication rate observed in most studies so far such as described Table 3-1.

3.1.6. The potential influence of deletions at 22q11.2 on DNA methylation

Deletions at chromosome 22q11.2 are the highest known schizophrenia risk factor identified to date, with about one third of patients with the deletion developing the disorder [57]. However, the deletion itself is not fully penetrant and suggests, as in idiopathic schizophrenia, the presence of other risk variants (Cf. Chapter 2 (p.31)), gene-gene interactions or gene-environment interactions. Epigenetic studies on twins, in particular monozygotic twins, is an elegant method to investigate such gene-environment interactions [286]. Despite sharing the same genotype, monozygotic twins have phenotypic differences, from minor traits to being discordant for certain disorders. It has been shown that during their lifetime, their epigenome accumulate changes, leading to significantly different transcriptomes [287]. Investigations of the methylome of monozygotic twins discordant for schizophrenia have revealed such epigenetic changes in genes and pathways relevant for the disorder [276,288,289]. Case reports of monozygotic twins with the 22q11.2DS have described discordant phenotypes, including congenital features (mainly congenital heart defects) [162–165,290,291]. One author reported different sizes of the deletion of 22q11.2 between two twins by using genomic DNA microarray analysis rather than the less precise techniques used in the other reports (fluorescence in situ hybridisation, qPCR). It suggests that this could be one reason for the phenotypic variability between them (non-psychiatric symptoms) [165]. As they share the same intrauterine environment, gene-environment interactions differences between twins would arguably be minimal

prenatally. It is therefore unlikely that any congenital defects that are discordant between a pair of twins with 22q11.2DS would be due to major environmental changes. On the other hand, neuropsychiatric symptoms in 22q11.2 appears later during lifetime, and are more likely to be affected by differences in DNA methylation.

No studies published so far have investigated genome-wide DNA methylation changes in 22q11.2 deletion patients, but certain genes within the 22q11.2 region are of particular interest in the context of methylation. The Catechol-O-Methyltransferase (*COMT*) gene, implicated in neurotransmitter inactivation, has been shown to have abnormal methylation in schizophrenia patients. In particular, in these patients the membrane-bound *COMT* (*MB-COMT*) gene promoter appears to be hypomethylated (in saliva [94] and brain tissue [95]) while the soluble *COMT* isoform (*S-COMT*) promoter is hypermethylated (in leukocytes [96]). Olanzapine, an atypical antipsychotic structurally similar to clozapine, has been shown to induce DNA methylation changes in rats cerebellum, hippocampus and liver [292]. Interestingly, when looking specifically at gene promoters inside the genomic region homologous to 22q11.2 in rats (59 genes), 34 showed altered methylation in response to treatment in at least one tissue type. These results suggest that changes in methylation and/or gene expression of genes within the deletion might influence the schizophrenia phenotype.

Two genes located within the common minimal 1.5mb deletion shared by all typical 22q11.2 deletion patients can potentially affect DNA methylation changes indirectly: *HIRA* and *DGCR8*. HIRA is a histone chaperone protein that has been shown to be essential for histone assembly and *de novo* DNA methylation in gametes/gametocytes [293,294]. The hemizygous deletion of *DGCR8*, involved in microRNA pre-processing (see Chapter 3 (p.49)), could potentially alter DNA methylation due to the transcriptome changes caused by abnormal microRNA levels. DGCR8 is part of the microprocessor complex acting in the canonical microRNA biogenesis pathway and its deletion leads to decreased microRNA levels [78,128]. It could affect mRNA and protein levels of genes involved in DNA methylation pathways; or it could affect transcription factors levels and thus change DNA methylation at certain genes promoters and enhancers (see Figure 3-1). In the former hypothesis, one could expect global and unspecific changes in DNA methylation, while in the latter it could lead to more targeted changes.

3.1.7. miRNAs and DNA Methylation

There is an interplay between microRNA and DNA methylation and that could be of importance in the context of the 22q11.2 deletion syndrome. First, the expression level of a microRNA is controlled by the promoter of the host gene or its precursor microRNA. Methylation levels at their respective promoter, enhancer regions, transcription start sites and/or gene bodies are therefore likely to affect the transcription of a microRNA. Accordingly, *Dnmt1/Dnmt3b* knock-out cell lines have shown that around 10% of all miRNAs tested were regulated by DNA methylation [295]. Haploinsufficient miRNAs spanned by the 22q11.2 deletion could be further affected by DNA methylation changes.

On the other hand, microRNAs have also been shown to mediate DNA methylation. *Dicer* (part of the microRNA biogenesis pathway, cf. Figure 4-1, Chapter 4 p.82) knockouts in embryonic stem cells (mESCs) have been shown to have a down-regulation of *de novo* DNA methyltransferase and methylation defects that can be rescued by the transfection of the miR-290 cluster miRNAs [296,297]. Similarly, *Dgcr8* mESCs knock-outs have reduced levels of *Dnmt3a*, *Dnmt3b* and *Dnmt3l* [298], with transfection experiments reporting that *Dnmt3b* and *Dnmt3l* levels are upregulated by miR-294 transfection and downregulated by let-7/miR-152 transfections [298,299]. Another microRNA, miR-185, has been shown to target DNMT1 in cancer lines, including gliomas [177,178]. It is of particular interest in the context of 22q11.2 deletion syndrome for two reasons. First, it is processed by DGCR8 and its level is decreased by around 20% the hippocampus of mice carrying a *Dgcr8* hemizygous deletion [176]. Secondly, miR-185 is located within the minimal 1.5Mb deletion at 22q11.2 and also in the syntenic region on mouse chromosome 16. A mouse model with a hemizygous deletion in this syntenic region (Df(16)$A^{+/-}$) has shown a 70 to 80% decrease of miR-185 in the hippocampus and prefrontal cortex [176], an effect potentially due to the hemizygous deletion of both miR-185 and *Dgcr8*.

A significant decrease of miR-185 levels can be found in blood samples of 22q11.2 deletion syndrome patients [129] and was associated with significant decrease in hippocampal volume. A previous study reporting that mice with conditional *Dnmt1* and/or *Dnmt3a* knock-out have smaller hippocampi volumes [300], which is also a feature observed in 22q11.2 deletion syndrome patients [301–303]. However, decreased levels of miR-185 should result in increased levels of *Dnmt1* and *Dnmt3a*/b, thus potentially not

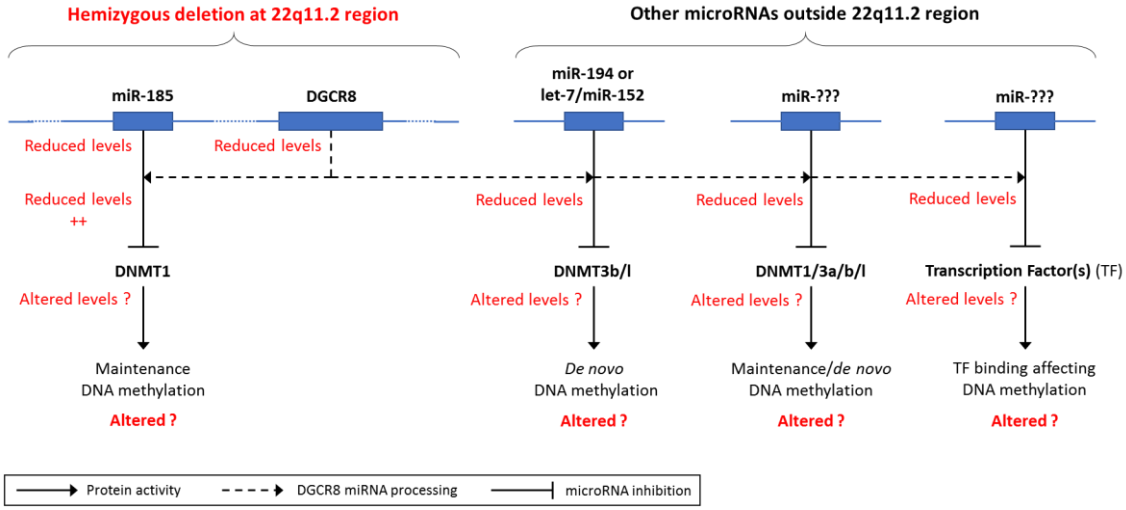


Figure 3-2. Study hypothesis: deletion at 22q11.2 region affects DNA methylation patterns through microRNA dysregulation. The hemizygous deletion of DGCR8 gene change microRNA levels, possibly resulting in altered levels of proteins that affects DNA methylation. The possible consequences of the 22q11.2 deletion are depicted in red.

having the similar effect than a knock-out of these genes. Finally, it has been demonstrated that miRNAs preferentially target genes with low DNA methylation levels in their promoter regions [304], suggesting that the two mechanisms might be complementary.

3.2. Aims of the chapter

In this chapter, I investigate DNA methylation profiles of 22q11.2 deletion syndrome patients. This is, to our knowledge, the first epigenome-wide association study of this disorder. Our hypothesis is that the dysregulation of microRNA expression through the hemizygous deletion of *DGCR8* and miR-185 will affect DNA methylation, potentially both in a direct way (dysregulation of methylation pathway enzymes levels) and indirect way (dysregulation of transcription factors levels). This hypothesis is summarised in Figure 3-2. The goals of the project are:

- Determine if DNA methylation is affected in 22q11.2DS patients

- Determine if genes that are differentially methylated and their related biological pathways have been previously shown to be associated with schizophrenia.

3.3. Methods

3.3.1. Sample origin

22q11.2DS samples analysed in this chapter are part of 2 independent cohorts:

- ECHO cohort: ExperienCes of people witH cOpy number variants cohort. The Cardiff ECHO study of children with copy number variations is led by Prof. Marianne van den Bree and Prof. Michael Owen at Cardiff University. Participants were referred by UK genetics clinics and charities. Blood samples from 22q11.2DS children and sibling controls were selected from the study [52].

- BBAG cohort: Brain, Behaviour And Genetics in 22q11.2 deletion syndrome cohort. The BBAG study is led by Dr. Michael Craig at King's College London. Blood samples provided were collected from 22q11.2DS patients and individual without the 22q11.2 deletion that were recruited for brain imaging studies principally.

3.3.2. Sample preparation

Genomic DNA was extracted from whole blood with PAXgene Blood DNA kits (Qiagen) and quantified with Quant-it PicoGreen dsDNA Assays (ThermoFisher Scientific). 500 ng of purified genomic DNA from each sample was treated with sodium bisulfite with the EZ-96 DNA Methylation-Gold Kit (Zymo Research). Genome-wide methylation of bisulfite-converted DNA was assessed using either Illumina InfiniumHumanMethylation450 BeadChip (Illumina) or Infinium HumanMethylationEPIC BeadChip (Illumina) and scanned on an Illumina HiScan system (Illumina). Illumina InfiniumHumanMethylation450 BeadChip (450K array) interrogate methylation levels at over 485000 sites per sample that cover 99% of RefSeq genes; Infinium HumanMethylationEPIC BeadChip (EPIC array) is an upgraded version with over 850000 methylation loci assessed, with around 93% of loci also contained on the 450K array (correlation between arrays types: $R^2 > 0.98$, information provided by Illumina). Genomic DNA was extracted by the MRC Centre for Neuropsychiatric Genetics and Genomics Core team. 450K arrays were prepared and run by Dr Afnan Salaka and Catherine Bresner (Cardiff University); EPIC arrays were prepared and run with the help of Catherine Bresner.

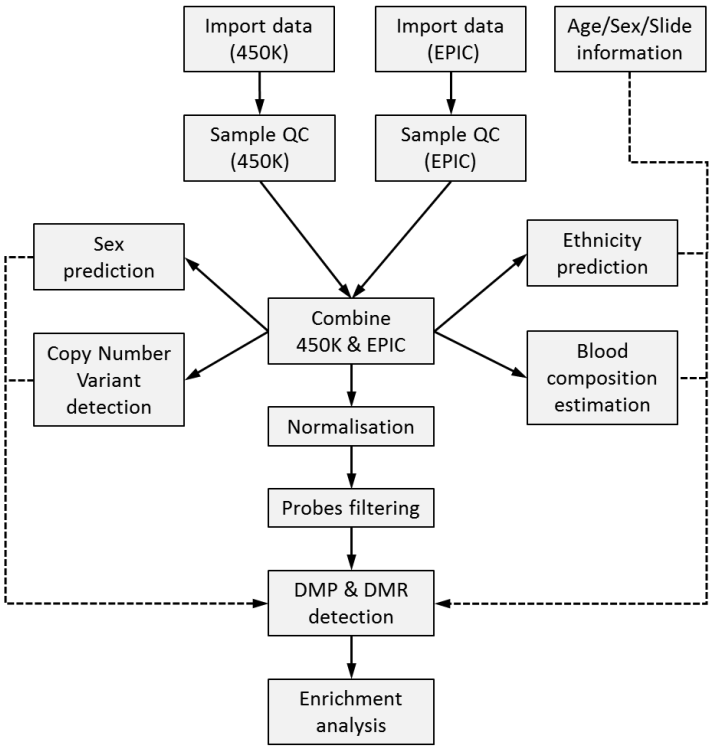


Figure 3-3. DNA methylation data processing pipeline. Dotted lines represent data used as covariate in the regression model fitted to detect DMP and DMR.450K: data from 450K arrays; EPIC: data from EPIC arrays; DMP/DMR: differentially methylated probes/regions; QC: quality control.

3.3.3. **Data processing and quality control**

The data processing workflow is presented in Figure 3-3. All data processing was performed in R (Microsoft R Open 3.4.0 (Microsoft), based on R-3.4.0 (R Statistics)). The *minfi* package and its various integrated functions were used to process the data unless specified otherwise. Data (IDAT files) from 450K and EPIC arrays were loaded into the R environment separately. Samples were included if they had less than 1% of sites with a detection p-value greater than 0.01, and if they passed the standard array Illumina quality controls (bisulphite conversion I and II controls, extension controls, hybridization controls, non-polymorphic controls, specificity I and II controls and target removal controls). The methylation levels at each probe was characterised by a Beta value, which is the ratio of the methylated probe intensity and the sum of methylated and unmethylated probes intensities. Subsequently, data from 450K and EPIC arrays were combined and normalised using the single-sample normalisation method ssNoob (derived from the Noob method [305]) as it has been shown to be the best method to date for joint analysis of data from 450K and EPIC arrays [306]. Quality control of combined data before

and after normalisation was assessed interactively with the *ChAMP* package (Beta value distributions shown in Supplementary figure 2-1). Low quality probes were then removed with the *wateRmelon* package (probes with at least 5% of samples with a beadcount < 3 and a detection p-value < 0.01). Probes located on the X and Y chromosomes, as well as previously recognised underperforming probes (such as cross-hybridising probes or probes containing a single-nucleotide polymorphism within 5 bp of the probe 3' end) were excluded from all analyses (Chen *et al.* [307], data available at http://www.sickkids.ca/MS-Office-Files/Research/WeksbergLab/48639-non-specific-probes-Illumina450k.xlsx and Zhou *et al.* [308], GRCh37/hg19 annotation with suggested overall masking, data available at http://zwdzwd.github.io/InfiniumAnnotation). After quality control and filtering, 381,688 probes were subsequently used for analysis.

3.3.4. Data-based phenotypic prediction and assessment of potential confounders

Different data-based analyses have been performed to assess possible confounders in the dataset as well as avoiding additional sample issues. For instance, it has been shown that sex and ethnicity affect global methylation levels in DNA extracted from blood [309].

- Methylation-based sex prediction was performed with the *minfi* package and compared to available phenotype information.

- Probes are present in the Illumina arrays to detect single nucleotides polymorphisms (SNPs) and allow the discrimination of samples (59 probes present in both 450K and EPIC arrays); these probes were used to detect (and remove) any duplicated samples.

- These "explicit" SNP probes have been previously used for race prediction and discriminate between African American and non-African American populations [273]. Moreover, a subset of CpG probes have been shown to act as "implicit" SNP probes because they contain a SNP within the probe that leads to a colour switch on the array depending on base change [308]. Joint use of "explicit" and "implicit" SNP probes have been shown to increase ethnicity prediction accuracy [308]. Of the 59 "explicit" and the top 30 most predictive (as described in the original paper) "implicit" probes, a total of 78 had beta values for all samples, and these were used to look for population stratification by principal component analysis (PCA). In order to increase discrimination by PCA, the data

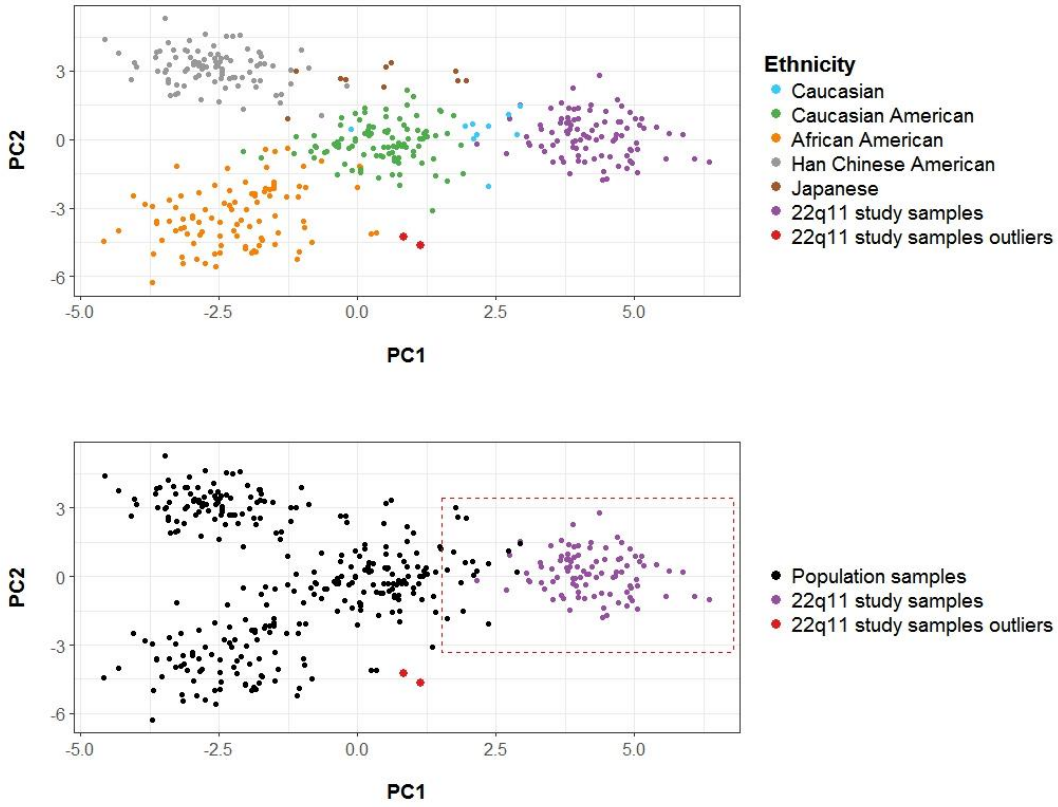


Figure 3-4. Population stratification and outlier detection. (A) Principal component analysis discrimination of samples from different ethnicity **(B)** Outlier detection based on standard deviation (3 standard deviation away from the means of both principal components, represented in red dotted lines).

was merged with publicly available DNA methylation from blood and lymphoblastoid cell lines from different ethnicity (GEO accession GSE36369: 96 Caucasian American, 96 Han Chinese American and 96 African American LCL samples; 10 Caucasian, 10 African American (GSE36064) and 10 Japanese blood samples.) (Figure 3-4.A). Outliers (samples further than 3 standard deviation away from the mean of the first two principal components) were excluded from the analysis (N=2) (Figure 3-4.B). The 4 first principal components were selected as covariates in the regression analysis, based on observation of PCA plots (Supplementary figure 3-3.A) and Scree plot (Supplementary figure 3-3.B), leading to a decrease in genomic inflation (Supplementary figure 3-3.C).

- Blood cell-type sample compositions were estimated with the *minfi estimateCellCounts* function and compared between 22q11.2DS cases and controls with a Bonferonni-corrected Wilcoxon test (cf. results and Figure 3-6). The cell proportions

(CD4+ T-cells, CD8+ T-cells, Monocytes, Granulocytes, Natural Killers and B-cells) were used as covariates in the regression analysis.

- Presence and size of deletion at locus 22q11.2DS have been assessed in all samples by copy number variant (CNV) detection with the *conumee* package, using as controls CNV-free samples from the *CopyNumber450kData* package (deprecated since 04/2017).

- Age outliers were removed (samples further than 4 standard deviation away from the mean age) (N=1) as well as samples without age information (N=1).

- CpG sites known to be associated with active smoking exposure in blood samples and/or prenatal tobacco smoke exposure have been removed. Active smoking exposure associated probes were selected from a systematic review of DNA methylation studies by Gao *et al*. [310] (14 EWASs and 3 gene-specific methylation studies, N = 1460 probes) and a more recent EWAS [311] (N = 748 probes). Prenatal smoking exposure associated probes are from a study in which the maternal plasma level of a biomarker of smoking has been assessed [312] (N = 26 probes passing Bonferonni-corrected significance; however the top 100 most significant probes were included) and a second study with self-reported maternal smoking during pregnancy [313] (N = 19). A total of 2058 probes associated with tobacco exposure have been removed prior to analysis; including these probes in the analysis shows that 27 of them are significantly differentially methylated between 22q11.2DS cases and controls.

The characteristics of sample groups are presented Figure 3-5.

3.3.5. Statistical analysis and enrichment analysis

Beta values were first transformed into M-values (M = log2(Beta / (1 – Beta))), as their distributional properties make them more suitable for the statistics used in the differential methylation analysis [314]. Detection of differentially methylated probes (DMPs) and regions (DMRs) was performed with the *DMRcate* package [315], based on the *limma* package that test association between case/control status and methylation levels with a linear regression model [316]. The model was fitted to the M-values with Age, Sex, Slide (also accounting for array type (450K or EPIC)), Cell type proportion estimations (6 cell types), Cohort (ECHO or BBAG) and Ethnicity-based principal components (4

Group	N	Age mean (s.d.)	Sex (F / M)	Array type (450k / EPIC)	Deletion type (None / Small / Large)	Cohort (ECHO / BBAG)
Controls	45	13.1 (5.6)	22 / 23	23 / 22	45 / 0 / 0	32 / 13
22q11.2del samples	42	14.6 (6.4)	17 / 25	21 / 21	0 / 3 / 39	23 / 19

Figure 3-5. Sample characteristics for cases (22q11.2 samples) and control groups. N: number of samples.

PC) as dependent variables. The model fit for each array has been determined with the *limma* function *arrayWeights* and the calculated weights were incorporated in the model. False discovery rate (FDR) was adjusted by Benjamini-Hochberg correction [317]. Detection of DMRs was done with *DMRcate* default settings (Lambda = 1000, C = 2, Benjamini-Hochberg correction for multiple testing, p-value cutoff of 0.05). Each DMR contained 2 consecutive CpG probes or more, and CpGs separated by more than Lambda (1000 bp) were in separate DMRs. FDR-corrected p-values were combined within each DMR with the Stouffer method [318] and only DMR with a Stouffer(FDR) < 0.05 were retained and described in the results.

KEGG and GO Enrichment analysis were done with the *gometh* function from the *missMethyl* package [319]. This function performs an over-representation analysis using a hypergeometric test and takes into account the number of probes per gene to adjust the probability of significant differential expression (based on the GOseq method [320]). Only CpG sites that passed filtering/quality control were used as the background set. The enrichment of DMPs within schizophrenia-GWAS associated regions was tested by Fisher's exact test. Prof. Jonathan Mill provided correlation data between blood and brain methylation levels, a description of which was reported in Hannon 2015 [321]. A probe was considered as having a blood/brain correlation if it had a Pearson correlation coefficient $r > 0.3$ (low/moderate correlation) between blood methylation levels and at least one brain region tested in the original study (prefrontal cortex, entorhinal cortex, superior temporal gyrus, and cerebellum).

Plots were been generated with the *ggplot2* package [242] unless specified otherwise.

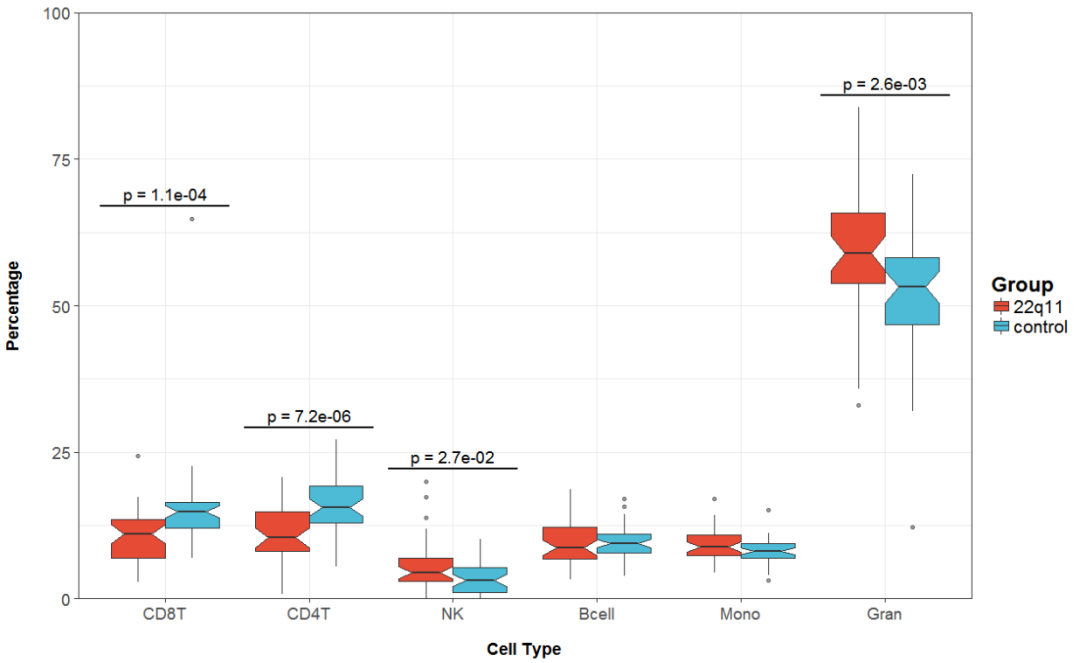


Figure 3-6. Estimation of leukocytes proportion in 22q11.2 deletion syndrome patients and controls. CD8T: CD8+ T-cell; CD4T: CD4+ T-cell; NK: Natural Killer cell; Bcell: B-cell; Mono: Monocyte; Gran: Granulocyte. Wilcoxon signed rank test, Bonferroni corrected for multiple comparisons.

3.4. Results

3.4.1. 22q11.2 samples altered cell-type composition lead to an important enrichment bias

An initial analysis of probes that were differentially methylated in 22q11DS relative to non-deleted controls revealed that 69.2% of all of CpG sites tested genome-wide were differentially methylated (*i.e.* 25542 differentially methylated probes (DMPs) with FDR < 0.05). After exclusions of the DMPs within the 22q11.2 region (cf. next result), a Gene Ontology enrichment analysis (biological process ontology) of the top 1000 most significantly DMPs revealed one significant enrichment after FDR correction: immune system process (FDR = $3.4*10^{-4}$). Analysis of these DMPs with eForge, a tool to identify cell- and/or tissue-specific signal in DNA methylation data [322], showed that the most significant tissue-specific enrichment was "foetal thymus" (Q-value $1.7*10^{-34}$, Supplementary figure 3-2.A).

Thymus aplasia is one of the most common feature of 22q11.2 deletion syndrome and is present in around 80% of patients [44]. This organ is essential for T-cell development [323],

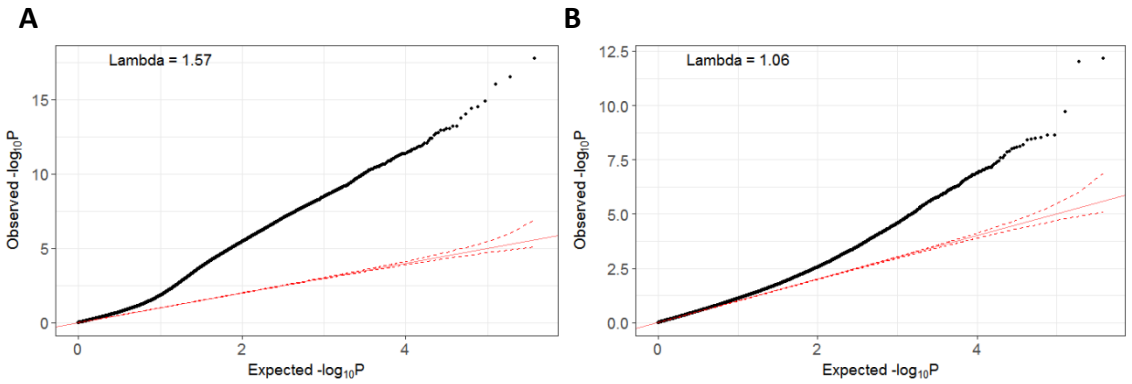


Figure 3-7. Correction of the genomic inflation due to differences in leukocytes proportion. **(A)** Before correction for cell type composition, **(B)** After correction. P-values from DMPs within the 22q11.2 region are excluded. A lambda close to 1 indicate that results are not inflated by unaccounted variance. Lambda = genomic inflation factor. The red line represents the uniform distribution; red dotted lines represent the 95% confidence interval.

and it has been shown that 22q11.2 patients have a significant deficit in this cell type, with differences decreasing during adulthood [43,44]. A deficit in one cell type in the patients will lead to difference in the proportion of leukocyte subtypes between 22q11DS patients and controls and could adversely result in the observed differential DNA methylation levels [324]. Koestler *et al.* have demonstrated in 2013 that it is possible to estimate the blood cell composition based on DNA methylation profiles [325] and studies have since demonstrated that adjusting for these estimates is essential [284,326].

The R package *minfi* was used to obtain leukocytes proportions in the 22q11DS samples and non-deleted controls, based on reference DNA methylation data from FACS-sorted adult blood [284]. The results revealed a deficit in T-cells in 22q11.2 deletion syndrome patients ($p = 2.2*10^{-5}$ (CD8T), $p= 1.0*10^{-5}$ (CD4T)). There was also an increase in Natural Killer cells ($p = 1.7*10^{-2}$) as well as Granulocytes ($p = 1.3*10^{-3}$) which might reflect a proportion change due to the T-cell deficits (Figure 3-6). While it has been suggested that this method based on adult reference data may not be appropriate in new-borns or young children [327], the mean age of the samples used in this study was 14.1, which is higher than the 12 years old patient group used in the study, for which the data correlate well with differential cell counts (ρ=0.75-0.77). Other methods for correcting for cell proportion heterogeneity have been proposed, however reference-based methods such as the one implemented in *minfi* appears to be the most appropriate when tissue-specific methylome data is available [282,328,329].

After adjustment for cell composition using cell type percentages as covariates in the regression model used to calculate differential methylation, only 662 DMPs remained significant. The quantitative-quantitative (Q-Q) plot in Figure 3-7 shows that adjusting for cell composition reduced the genomic inflation from Lambda = 1.57 to 1.06. Gene Ontology enrichment analysis of this adjusted data did not show significant enrichment in immune-related biological processes. Moreover, eForge analysis of the top 1000 most significant DMPs showed an important decrease in the significance of the enrichment in thymus and blood-specific CpGs sites (Supplementary figure 3-2.B).

3.4.2. Hemizygous deletion of a genomic region leads to potential spurious associations

Differentially methylated probes (DMPs) and differentially methylated regions (DMRs) were calculated after adjusting for cell composition. Analysis of the distribution revealed an important association of probes located within the 22q11.2 region (Figure 3-8). In fact, around 42% of the probes in this region were differentially methylated, compared to only 0.2% of all other probes outside of the 22q11 deleted region. The window of differentially methylated probes on chromosome 22 is clearly delimited by the boundaries of the deletion (Figure 3-8.C.3 and Figure 3-8.C.4). In order to try and determine if these differences were due to a technical artefact or a true biological signal a number of tests were then conducted.

3.4.3. Investigating the source of the increased rate of differential DNA methylation at 22q11

If the differences were an artefact due to reduced probes intensities caused by hemizygosity at 22q11 in deletion carriers, it could be expected to result in a similar direction of effect for all affected probes. However, as can be seen in Figure 3-8.A.2 the DMPs spanning the 22q11 deletion can be either hypo- or hyper- methylated in 22q11DS patients relative to non-deleted controls. On the other hand, Figure 3-8.B.1 and Figure 3-8.B.2 shows that DMPs and DMRs within this region have very low p-values, despite small differences in beta-values. It is therefore possible that hemizygosity has resulted in an inflation of the p-values for probes spanning the 22q11 deletion.

In order to establish whether hemizygosity resulted in falsely inflated evidence for differential methylation, the data were compared to the only independent but comparable

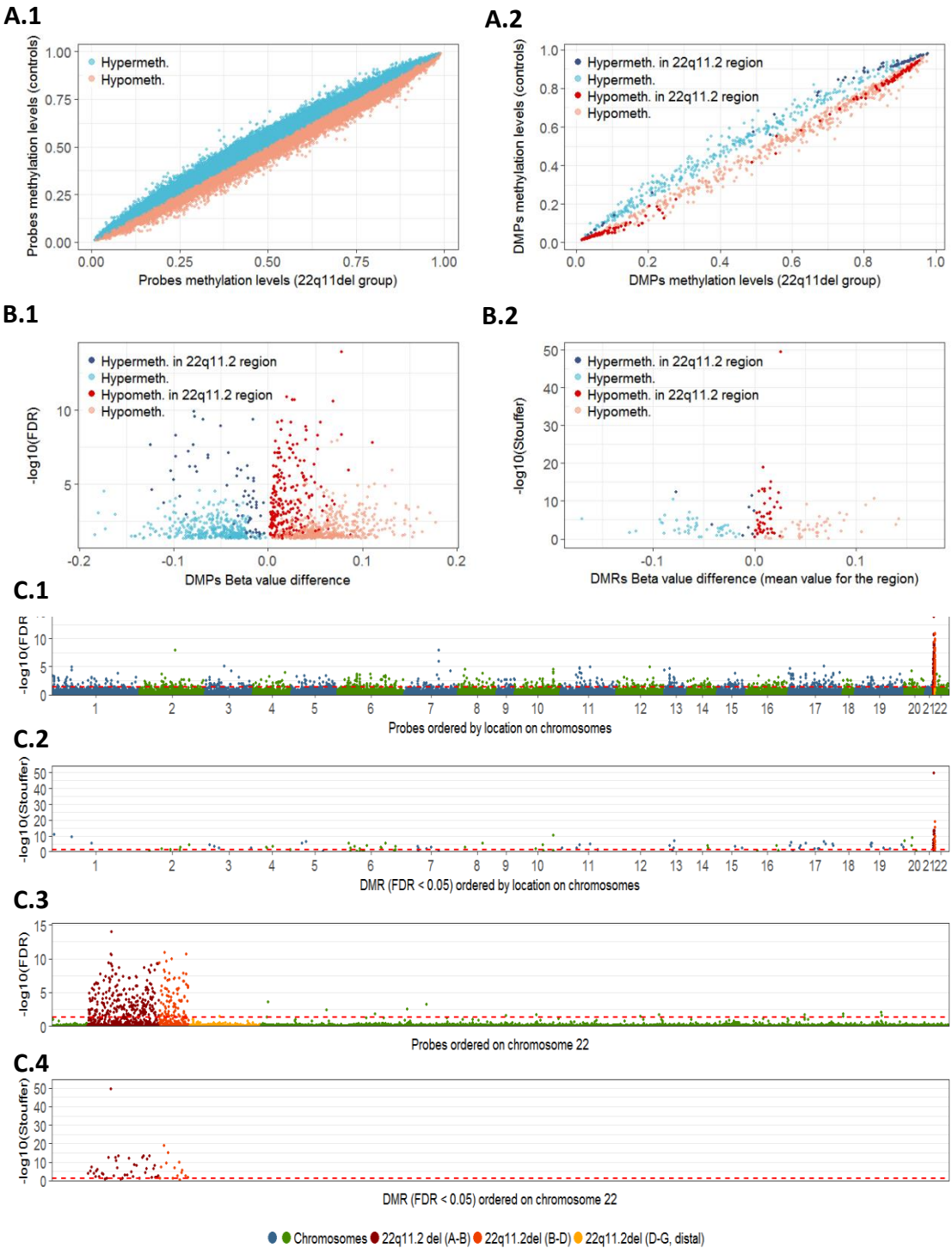


Figure 3-8. Overview of differential methylation results and specific association at the 22q11.2 region. (A) Comparison of methylation levels between cases (22q11.2 group) and controls for all **(A.1)** or only significant **(A.2)** probes. **(B)** volcano plots of DMP **(B.1)** and DMR **(B.2)**. **(C)** Manhattan plots for all probes **(C.1)** and DMRs **(C.2)** ordered by chromosomal location, with a zoom-in on chromosome 22 (**C.3** and **C.4**). The red dotted lines symbolize significance level (FDR < 0.05 or Stouffer < 0.05). 22q11.2 deletion breakpoints LCR A-B are the boundaries of the small 1.5 mb deletion; the large 3Mb deletion is between breakpoints LCR A-D, and the distal deletion is between breakpoints D-G. Hypermeth/hypometh: increased/decreased methylation in 22q11.2 patients; FDR: false discovery rate (Benjamini–Hochberg procedure); Stouffer: Stouffer transformation of the group of FDRs value for CpGs part of DMRs.

CNV case-control study available to our knowledge, from Strong *et al.*[330] (GEO accession

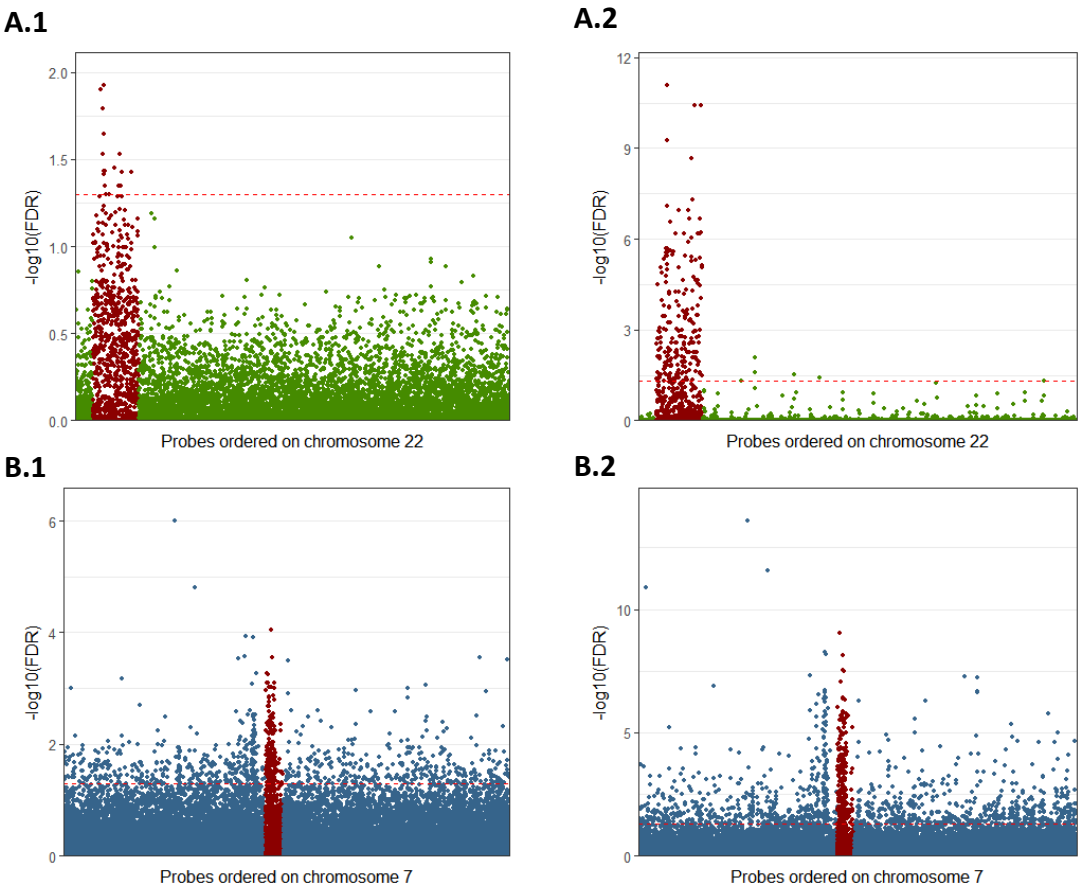


Figure 3-9. Manhattan plots of regional associations at hemizygously deleted locus for different sample size. (A.1) Association for probes on chromosome 22 for N=20 22q11.2DS cases and N=15 controls. **(A.2)** Association for probes on chromosome 22 for N=20 22q11.2DS cases and N=66 controls. **(B.1)** Association for probes on chromosome 7 for N=20 WS cases and N=15 controls. **(B.2)** Association for probes on chromosome 7 for N=20 WS cases and N=66 controls. 7q11.3 deletion data is from GEO accession GSE66552. Probes within the hemizygously deleted region are in red. For (A.2) and (B.2), controls are a combination of controls from both studies. The red dotted lines symbolize significance level (FDR < 0.05). FDR: false discovery rate (Benjamini–Hochberg procedure).

GSE66552). This publicly available 450k methylation data included 15 control samples and 20 samples with Williams syndrome (WS). WS is a neurodevelopmental disorder caused by a 1.6mb hemizygous deletion on chromosomal region 7q11.23, causing various symptoms including cardiovascular abnormalities, characteristic facial features, growth retardation as well as cognitive and neuropsychiatric symptoms [331]. The study data processing and analysis was relatively similar to that conducted for 22q11DS in this thesis, but no particular regional association have been reported for the 7q11.23 locus in the published study.

However, analysis of the publicly available DNA methylation data of Strong *et al.* using the Stouffer method [318] which combined the FDR of all CpGs within the 7q11.23

deletion locus, revealed a highly significant excess of differential methylation in deletion carriers relative to non-deleted controls (Stouffer(FDR) = $1.62*10^{-50}$). In comparison, randomly selecting the same sample size in the 22q11DS/non-deleted control cohort (20 22q11.2DS samples + 15 controls) did not show significant evidence for association (Stouffer(FDR) = 1). It is possible that this difference could be due to differences in the ratio signal/background between the two studies, which is likely to reduce the power of the 22q11DS study relative to the WS study.

To address this, the sample size in both studies was increased by combining the control samples of both studies (N=66) and keeping the same number of cases (either 20 WS samples or 20 22q11.2DS samples). Analysis of these cohorts revealed that when probes that were spanned by either a 7q11.23 or a 22q11.2 deletion were tested for DNA methylation, highly significant evidence for differential methylation (relative to non-deleted controls) was obtained (Stouffer(FDR) = $1.23*10^{-160}$ (WS), Stouffer(FDR) = $4.26*10^{-39}$ (22q11.2DS)). Manhattan plots of the chromosomes containing these regions in these different scenarios are represented Figure 3-9. Random permutations over either the same number of CpGs than those included within the deletions or any CpGs included in genomic regions of same size than the deletions confirmed this significant association of these regions (p = 0; 10,000 permutations comparing Stouffer(FDR) values obtained through random sampling with the values obtained for N=66 controls and N = 20 cases in both cases).

Further analyses of differential DNA methylation in sufficiently powered cohorts for different CNVs will be required to further confirm that the observation is a technical artefact caused by hemizygosity. However, as the available data is not able to rule out the possibility of this being an artefact and in order to be conservative, all probes contained within the 22q11.2 deletion locus were excluded from any of the following differential methylation analyses.

3.4.4. Differential methylation analysis (DMP/DMR)

After excluding probes within the 22q11.2 region and correcting for known bias, the comparison of DNA methylation between 22q11.2DS cases and non-deleted controls identified 662 significant differentially methylated probes (DMPs, FDR < 0.05, 357 and 497 probes hypo- and hyper-methylated in 22q11.2 samples respectively) and 62

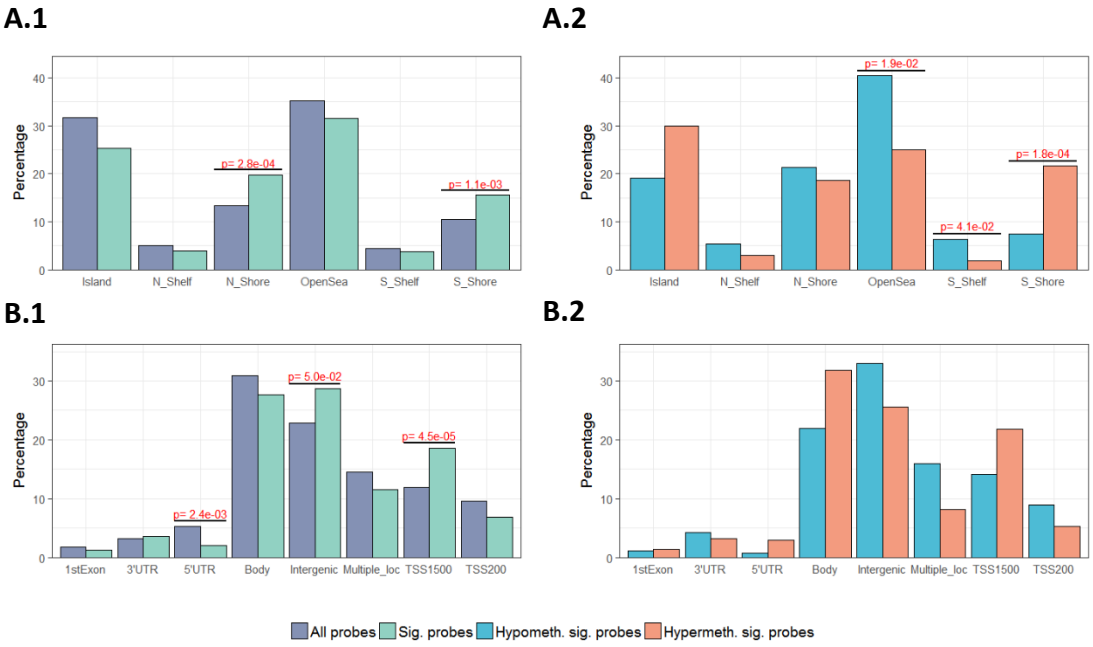


Figure 3-10. Differentially methylated probes (DMPs) features. (A.1/2) Relation to CpG islands for all probes versus significant probes **(A.1)** or significantly hypomethylated probes versus significantly hypermethylated probes **(A.2)**. **(B.1/2)** Genomic location for all probes versus significant probes **(B.1)** or significantly hypomethylated probes versus significantly hypermethylated probes **(B.2)**. A significant probe is a probe with a FDR-corrected p-value < 0.05. Island: CpG island (CGI); N_Shore/S_Shore: North/South CGI shores (sites located within +/- 2 kb from a CGI); N_Shelf/S_Shelf: North/South CGI shelves (sites located within +/- 2 kb from a CGI shore); North/South: upstream/downstream; OpenSea: sites not located within an annotated CGI location; UTR: untranslated region; TSS200/1500: regions located 200/1500 bp upstream of a transcription start site; Multiple_loc: probes located within multiple annotated genomic regions. Cgi-square test, bonferonni corrected for the multiple features tested. Functional annotation from the Illumina HM450 annotation (with Multiple_loc feature added).

significant differentially methylated regions (DMRs including at least 2 DMPs, Stouffer(DMR) < 0.05, 32 and 30 regions hypo- and hyper-methylated in 22q11.2 samples respectively). 66 significantly methylated probes were located at a CpG that has been shown previously to have methylation levels that are correlated between blood and brain tissue [321]. The DMPs were further broken down according to their features (Figure 3-10). This revealed that DMPs were significantly enriched in CGI shores ($p = 2.8*10^{-4}$ (North shores), $p = 1.1*10^{-3}$ (South shores)). There was also an enrichment in DMPs in the promoter regions located within 1.5kb of the transcription start site ($p = 4.5*10^{-5}$) but a deficit in the 5'UTR regions ($p = 2.4*10^{-3}$).

A

Pathway	N	DE	FDR	Pathway	N	DE	FDR
Metabolic pathways	1186	23	8.3e-07	Phagosome	139	5	1.3e-02
PI3K-Akt signaling pathway	317	12	3.5e-05	Cell adhesion molecules (CAMs)	131	5	1.6e-02
ECM-receptor interaction	79	7	5.7e-05	Malaria	46	3	2.1e-02
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	69	6	4.1e-04	Regulation of actin cytoskeleton	199	6	2.2e-02
Alcoholism	163	7	7.0e-04	Proteoglycans in cancer	198	6	2.6e-02
EGFR tyrosine kinase inhibitor resistance	77	6	7.0e-04	Fatty acid metabolism	46	3	3.6e-02
Focal adhesion	190	8	1.1e-03	Pathways in cancer	383	8	3.6e-02
Systemic lupus erythematosus	116	5	2.5e-03	Lysosome	117	4	4.3e-02
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	19	3	8.2e-03	Hippo signaling pathway	151	5	4.3e-02
Ras signaling pathway	217	7	8.2e-03	Endocytosis	251	6	4.3e-02
Adipocytokine signaling pathway	66	4	1.3e-02	MAPK signaling pathway	244	6	4.9e-02

B

Pathway	N	DE	FDR
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	69	5	1.4e-03
Metabolic pathways	1186	13	1.4e-03
ECM-receptor interaction	79	4	2.0e-02
Hippo signaling pathway	151	5	2.1e-02
Phagosome	139	4	3.2e-02
PI3K-Akt signaling pathway	317	6	3.2e-02
Cell adhesion molecules (CAMs)	131	4	3.3e-02
Insulin signaling pathway	131	4	3.3e-02
Hematopoietic cell lineage	88	3	3.4e-02
Herpes simplex infection	163	4	3.4e-02
Endometrial cancer	48	3	3.6e-02

Figure 3-11. KEGG pathway enrichment analysis. (A) DMPs with FDR < 0.05 (N_{DMP} = 662). **(B)** DMPs with FDR < 0.05 and |ΔBeta| > 0.05 (N_{DMP} = 401). N: number of gene in the pathway; N_{DMP}: number of DMP tested; DE: number of genes associated with at least one DMP; FDR: false discovery rate (the number of probes per gene is taken into account to adjust the probability of significant differential expression). Only significant KEGG pathway enrichment (FDR < 0.05) are shown.

3.4.5. Pathway analysis

Performing KEGG pathway enrichment analysis for all significant DMPs (N=662, FDR < 0.05, adjusting for the number of probes per genes) revealed significant evidence that 22 different biological pathways were enriched (Figure 3-11.A). Restricting this analysis to the DMPs with a difference in beta values |ΔBeta| > 0.05 (5% difference in methylation levels between cases and controls, N=477) showed an enrichment in 11 different pathways (Figure 3-11.B). The gene overlaps between pathways shown in Figure 3-11.A are represented in Supplementary figure 3-4 and Supplementary figure 3-5. Despite the majority of DMPs being hypermethylated rather than hypomethylated (380 vs 281 respectively), the enriched pathways were due primarily to the DMPs that were less methylated in 22q11.2 patients relative to controls (Supplementary figure 3-6). Restricting to DMPs that have a correlated methylation levels between blood and brain tissue did not show significant KEGG pathways enrichments (data not shown).

A

Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain	Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain
chr1:1564422-1565541	1120	7	2.2e-09	0.12	MMP23B, MIB2	TRUE	chr19:12958803-12958913	111	3	1.1e-03	0.04	MAST1	FALSE
chr10:129533731-129534613	883	6	8.4e-09	-0.08	AL391005.1, FOXI2	TRUE	chr15:74218418-74219645	1228	12	1.2e-03	-0.04	LOXL1, LOXL1-AS1	FALSE
chr17:1961109-1961778	670	5	1.4e-07	0.08	HIC1	FALSE	chr17:17109239-17110353	1115	13	1.5e-03	0.07	PLD6	FALSE
chr20:748992-749620	629	8	1.7e-06	-0.06	SLC52A3	FALSE	chr20:30618874-30619244	371	5	1.5e-03	0.07	RP1-310O13.7	FALSE
chr13:52769289-52769951	663	5	2.8e-06	0.08	MRPS31P5	TRUE	chr17:76354621-76355288	668	4	1.9e-03	-0.03	SOCS3, RP11-806H10.4	FALSE
chr17:59554672-59554746	75	2	9.1e-06	0.09	TBX4	FALSE	chr4:77723057-77723415	359	4	2.0e-03	0.06	RP11-359D14.3	FALSE
chr6:106546704-106546824	121	3	2.6e-05	-0.07	PRDM1, RP1E15.3	FALSE	chr3:42201314-42201898	585	6	2.4e-03	-0.06	TRAK1	FALSE
chr8:101348456-101348501	46	3	3.6e-05	-0.06	RNF19A, KB1G8.1	FALSE	chr6:152011656-152011666	11	2	2.4e-03	-0.03	ESR1	FALSE
chr5:38445220-38446192	973	9	4.7e-05	-0.07	EGFLAM	FALSE	chr13:36048892-36051749	2858	17	3.1e-03	-0.03	MAB21L1, NBEA	FALSE
chr6:28226885-28227482	598	11	6.1e-05	0.07	NKAPL, ZKSCAN4	TRUE	chr11:32451777-32452771	995	8	3.7e-03	0.05	WT1	FALSE
chr1:61517807-61517876	70	2	6.1e-05	-0.17	NFIA	FALSE	chr7:81399308-81399747	440	6	4.9e-03	-0.04	HGF	FALSE
chr19:12984170-12984645	476	2	6.3e-05	0.14	AC020934.1, MAST1, HOOK2	TRUE	chr4:38869552-38869779	228	5	6.0e-03	0.04	FAM114A1, MIR574	FALSE
chr19:13120987-13121571	585	2	8.0e-05	-0.09	NFIX	FALSE	chr6:74063982-74064594	613	7	7.3e-03	-0.07	DPPA5	TRUE
chr17:63554468-63554536	69	2	1.1e-04	0.09	AXIN2	FALSE	chr17:14206572-14207530	959	10	8.4e-03	-0.06	RP11O1.2	FALSE
chr19:56215104-56215575	472	4	2.2e-04	0.07	AC010525.5	TRUE	chr17:46680288-46681401	1114	8	1.2e-02	-0.08	HOXB6, HOXB-AS3, HOXB3	FALSE
chr11:65314454-65315625	1172	8	2.5e-04	0.05	LTBP3	FALSE	chr19:37958707-37959104	398	3	1.2e-02	0.06	ZNF569, ZNF570	TRUE
chr19:12305553-12306198	646	4	2.7e-04	0.07	CTD6L21.1	FALSE	chr4:186732837-186733060	224	7	1.3e-02	-0.09	SORBS2	FALSE
chr6:33043841-33044345	505	4	5.2e-04	0.06	HLA-DPB1	FALSE	chr15:90543224-90544015	792	6	1.3e-02	-0.03	ZNF710	FALSE
chr16:75468281-75468318	38	3	5.5e-04	0.14	CFDP1	TRUE	chr11:705692-705748	57	2	1.4e-02	0.05	EPS8L2, DEAF1	FALSE
chr3:16216094-16216606	513	5	5.8e-04	-0.06	GALNT15	FALSE	chr3:49170599-49171343	745	6	1.6e-02	-0.08	LAMB2	FALSE
chr11:45376874-45377324	451	5	8.3e-04	0.04	RP11-430H10.1	FALSE	chr6:76203225-76203675	451	7	1.7e-02	0.07	FILIP1	FALSE
chr17:1466709-1466715	7	3	8.3e-04	0.04	PITPNA	FALSE	chr7:27142100-27144595	2496	25	2.0e-02	-0.05	HOXA2	FALSE
chr14:94462309-94462607	299	4	9.3e-04	-0.06	LINC00521	FALSE	chr19:58446312-58447434	1123	11	2.4e-02	0.03	ZNF418	FALSE
chr6:30458519-30458730	212	4	1.1e-03	0.05	HLA-E	FALSE	chr6:31690725-31692375	1651	31	4.8e-02	0.03	C6orf25, LY6G6C	FALSE

B

Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain
chr1:1564422-1565541	1120	7	2.2e-09	0.12	MMP23B, MIB2	TRUE
chr10:129533731-129534613	883	6	8.4e-09	-0.08	AL391005.1, FOXI2	TRUE
chr17:1961109-1961778	670	5	1.4e-07	0.08	HIC1	FALSE
chr20:748992-749620	629	8	1.7e-06	-0.06	SLC52A3	FALSE
chr13:52769289-52769951	663	5	2.8e-06	0.08	MRPS31P5	TRUE
chr17:59554672-59554746	75	2	9.1e-06	0.09	TBX4	FALSE
chr6:106546704-106546824	121	3	2.6e-05	-0.07	PRDM1, RP1E15.3	FALSE
chr8:101348456-101348501	46	3	3.6e-05	-0.06	RNF19A, KB1G8.1	FALSE
chr5:38445220-38446192	973	9	4.7e-05	-0.07	EGFLAM	FALSE
chr6:28226885-28227482	598	11	6.1e-05	0.07	NKAPL, ZKSCAN4	TRUE

C

Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain
chr6:31690725-31692375	1651	31	4.8e-02	0.03	C6orf25, LY6G6C	FALSE
chr7:27142100-27144595	2496	25	2.0e-02	-0.05	HOXA2	FALSE
chr13:36048892-36051749	2858	17	3.1e-03	-0.03	MAB21L1, NBEA	FALSE
chr17:17109239-17110353	1115	13	1.5e-03	0.07	PLD6	FALSE
chr15:74218418-74219645	1228	12	1.2e-03	-0.04	LOXL1, LOXL1-AS1	FALSE
chr6:28226885-28227482	598	11	6.1e-05	0.07	NKAPL, ZKSCAN4	TRUE
chr19:58446312-58447434	1123	11	2.4e-02	0.03	ZNF418	FALSE
chr17:14206572-14207530	959	10	8.4e-03	-0.06	RP11O1.2	FALSE
chr5:38445220-38446192	973	9	4.7e-05	-0.07	EGFLAM	FALSE
chr20:748992-749620	629	8	1.7e-06	-0.06	SLC52A3	FALSE

D

Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain
chr1:61517807-61517876	70	2	6.1e-05	-0.17	NFIA	FALSE
chr19:13120987-13121571	585	2	8.0e-05	-0.09	NFIX	FALSE
chr4:186732837-186733060	224	7	1.3e-02	-0.09	SORBS2	FALSE
chr10:129533731-129534613	883	6	8.4e-09	-0.08	AL391005.1, FOXI2	TRUE
chr17:46680288-46681401	1114	8	1.2e-02	-0.08	HOXB6, HOXB-AS3, HOXB3	FALSE
chr3:49170599-49171343	745	6	1.6e-02	-0.08	LAMB2	FALSE
chr6:106546704-106546824	121	3	2.6e-05	-0.07	PRDM1, RP1E15.3	FALSE
chr5:38445220-38446192	973	9	4.7e-05	-0.07	EGFLAM	FALSE
chr6:74063982-74064594	613	7	7.3e-03	-0.07	DPPA5	TRUE
chr20:748992-749620	629	8	1.7e-06	-0.06	SLC52A3	FALSE

E

Location	Width	# CpGs	Stouffer	deltaBeta	Genes	BloodBrain
chr19:12984170-12984645	476	2	6.3e-05	0.14	AC020934.1, MAST1, HOOK2	TRUE
chr16:75468281-75468318	38	3	5.5e-04	0.14	CFDP1	TRUE
chr1:1564422-1565541	1120	7	2.2e-09	0.12	MMP23B, MIB2	TRUE
chr17:59554672-59554746	75	2	9.1e-06	0.09	TBX4	FALSE
chr17:63554468-63554536	69	2	1.1e-04	0.09	AXIN2	FALSE
chr17:1961109-1961778	670	5	1.4e-07	0.08	HIC1	FALSE
chr13:52769289-52769951	663	5	2.8e-06	0.08	MRPS31P5	TRUE
chr6:28226885-28227482	598	11	6.1e-05	0.07	NKAPL, ZKSCAN4	TRUE
chr19:56215104-56215575	472	4	2.2e-04	0.07	AC010525.5	TRUE
chr19:12305553-12306198	646	4	2.7e-04	0.07	CTD6L21.1	FALSE

Figure 3-12. Differentially methylated regions (DMRs) associated with a promoter region. **(A)** All DMRs, sorted by significance. **(B)** Top 10 most significant DMRs. **(C)** Top 10 DMRs with most significant probes. **(D)** Top 10 most hypomethylated DMRs. **(E)** Top 10 most hypermethylated DMRs. BloodBrain: presence of probes that have shown to be correlated between blood and brain tissue. DeltaBeta: average beta value differences for the probes (deltaBeta > 0: hypermethylation in 22q11.2DS samples); Genes: genes with a promoter region that contain the DMR; Location: genomic location. Stouffer: combined false discovery rates for the probes, with the Stouffer method; Width: DMR size.

3.4.6. Differentially methylated regions

Of the 62 differentially methylated regions, 57 were associated with the promoter region of an annotated gene (Figure 3-12.A). The most hypomethylated and hypermethylated regions are presented Figure 3-12.C and Figure 3-12.D. Changes in DNA methylation ranged from a 14% increase for the most hypermethylated region

CpG ID	Location	deltaBeta	FDR	Associated Gene(s)	Distance TSS	CGI	Gene region	Enhancer region	BloodBrain
cg10316764	chr16:29984665-29984665	0.057	3.5e-02	TAOK2	521	Island	TSS1500	FALSE	FALSE
cg21045306	chr12:57637925-57637925	0.060	3.7e-02	STAC3	3449	S_Shore	Body	FALSE	TRUE

Figure 3-13. DMPs linked to genes associated with schizophrenia in GWAS. DMPs overlapping a LD-independent genome-wide significant SNP associations for schizophrenia (N= 108 regions tested). Location: Genomic location; deltaBeta: difference in beta values between controls and 22q11.2DS samples (deltaBeta > 0: hypermethylation in 22q11.2DS samples); FDR: False Discovery Rate; Dist.Prom: distance to the transcription start site (TSS) of the nearest gene; CGI: relation with CpG Islands; BloodBrain: probes that have shown to be correlated between blood and brain tissue.

(included within *AC020934.1*, *MAST1* and *HOOK2* genes promoter regions) to a 17% decrease for the most hypomethylated region (within *NFIA* promoter region). These regions included between 2 and 31 CpGs (for the DMR within the *c6orf24* and *LY6G6C* gene promoter regions). 9 of the DMRs include at least one DMP for which methylation levels have been previously shown to be correlated between blood and brain tissue [321]. The DMRs that were not associated with a gene promoter region are presented in Supplementary figure 3-7.

3.4.7. DMP enrichment in schizophrenia-related genes

A recent large-scale genome-wide association study has detected 128 genomic regions associated with schizophrenia [92]. Some of these associations are overlapping, thus there are 108 different loci in total that are associated with the disorder. 2 DMPs were located within these regions (Figure 3-13.A), which was less than expected by chance ($p = 0.029$, Fisher's exact test). The *STAC3*-associated DMP has been previously to have a correlation between DNA methylation in blood and in brain tissue.

3.5. Discussion

Comparison of DNA methylation between 22q11.2 deletion syndrome patients and controls revealed 662 significantly differentially methylated probes and 62 significantly differentially methylated regions containing at least two probes. DMPs were significantly enriched in CpG Islands shores ($p = 2.8*10^{-4}$ (North shores), $p = 1.1*10^{-3}$ (South shores)). This pattern may reflect differences in tissue/cell types but CGI shores are also strongly associated with gene expression [332]. Furthermore, there was also an enrichment of

methylation in the promoter region (located within 1500bp from the transcription start site, $p = 4.5*10^{-5}$) which can be negatively correlated with gene expression [249].

This study compared individuals with or without the 22q11.2 deletion, without controlling for individual phenotypes. This meant that it was not possible to test for particular pathways that were associated with specific phenotypes encountered in 22q11.2 deletion syndrome patients. Nevertheless, it did generate data that can potentially help gain a better understanding of the biology behind this disorder, in particularly those with a high incidence in 22q11.2, such as schizophrenia. Differential methylation analysis of all DMPs (regardless of difference levels) implicated 22 KEGG pathways, 15 of which could be grouped into 4 main (overlapping) groups with a biological relevance to schizophrenia: metabolism, cell signalling, cell adhesion and intracellular vesicle trafficking (1: Metabolism: metabolic pathways, fatty acid metabolism, adipocytokine signalling pathway, insulin signalling pathway; 2: Cell adhesion: ECM-receptor interaction, focal adhesion, cell adhesion molecules (CAMs); 3: Vesicle trafficking: endocytosis, lysosome, phagosome, regulation of actin cytoskeleton; 4: Cell signalling: PI3K-Akt signalling pathway, MAPK signalling pathway, Ras signalling pathway, hippo signalling pathway, EGFR tyrosine kinase inhibitor resistance, insulin signalling pathway, adipocytokine signalling pathway). It is important to note that many of these pathways share common genes with other pathways within and between these 4 groups, as shown in Supplementary figure 3-4 and Supplementary figure 3-5.

Early transcriptome analyses of schizophrenia brain compared to matched controls suggested that multiple metabolic pathways were affected in these patients [333], particularly energy metabolism which was later shown to be affected through mitochondria dysfunction [334]. Glucose metabolism has been shown to be affected, as well as fatty acid metabolism, in response to the resulting oxidative stress [334]. Metabolic changes have also been observed in 22q11.2DS patients, including increased fatty acid plasma concentration that has been suggested to be due to the haploinsufficiency of the mitochondrial citrate transporter *SLC25A1* [335]. Moreover, another gene within the deletion, *ZDHHC8*, codes for a palmitoyltransferase. Palmitoylation is a posttranslational process that add fatty acids chains to proteins. Around 40% of synaptic proteins are palmitoylated [336] and ZDHHC8 function seems to be essential for brain and neuronal development [122]. A study has shown that fatty acids such as long-chain ω-3 (omega-3)

polyunsaturated fatty acids could reduce the risk of transition to schizophrenia in high-risk individuals [337], however this has not been replicated in a larger multi-center trial [338].

Cell adhesion molecules (CAMs) are essential for synaptogenesis (as reviewed in Washbourne 2004 [339]), and synaptic dysfunction is suggested to have an important role in schizophrenia [340]. The extra-cellular matrix (ECM) interact with the cells during this process, and it is also important for cell adhesion and migration which are altered in schizophrenia [341]. In particular, a class of ECM molecules named proteoglycans (such as heparan and chondroitin sulfate proteoglycans) have been linked to neuron migration and associated with schizophrenia (as reviewed in Maeda 2015 [342]). Notably, the 22q11.2-deleted gene *DGCR2* is a putative adhesion receptor protein, and a murine model of 22q11.2DS has been reported to present defects in interneuron migration [343].

It has been suggested that endocytosis and protein trafficking are pathophysiological processes in schizophrenia, and that antipsychotic drugs interact with proteins involved in endocytosis [344]. Moreover, transcriptomic and genome-wide association study analyses have showed an enrichment in schizophrenia patients of genes involved in a network centred on the regulation of actin cytoskeleton and lysosomal function, implying that intracellular vesicle trafficking is also relevant to this disease [345]. In the context of 22q11.2DS, it has been shown that miR-185, whose levels are decreased due to its presence in the deletion and its processing by DGCR8, have two validated targets (*RhoA* and *Cdc42*) involved in the regulation of actin cytoskeleton (as reviewed by Forstner 2013 [346]).

The PI3K-AKT-GSK3-mTor and the Ras-ERK (/Ras-MAPK) signalling pathways are two intracellular signalling pathways essential to cell survival and involved in many processes such as cell proliferation and differentiation. Both pathways are cross-regulated in a system of feedback circuits [347]. HIPPO signalling pathway, which is significantly enriched in this study, has also been shown to be part of this complex signalling system and is indirectly regulated by EGFR-Ras-MAPK signalling [348]. Due to their primordial functions, defects in these pathways have been involved in many disorders such as cancer but also schizophrenia [349,350]. Analysis of DNA methylation differences between monozygotic twins discordant for schizophrenia has also shown an enrichment of differentially methylated genes involved in both MAPK and HIPPO signalling pathways [289]. CRKL is a protein whose gene is located at the 22q11.21 locus that mediates MAPK

signalling. Mouse knock-out models of CRKL have been shown to recapitulate some phenotypes observed in 22q11.2DS models [351], suggesting the importance of this pathway in this syndrome. It should however be noted that CRKL is located in a part of the deletion not shared by every patient. *ZDHHC8* on the other hand is within the A-B deleted region at 22q11.2 and its protein has an effect on the AKT/GSK3 pathway through palmytoilation of CDC42 [121]. Both single gene mouse knock-out models of *ZDHHC8* and a murine model of 22q11.2DS have been reported to result in phenotypes such as abnormal neuromorphogenesis and behaviour that can be rescued by downregulating GSK3 during development [121,352].

Finally, GSK3 have been shown to be involved in arrhythmogenic cardiomyopathy [353], one of the KEGG pathway enriched in the present analysis. Moreover, *Gsk3*$^{-/-}$ knock-out mice exhibit different cardiac defects[354], including ventricular septal defects, which are common in 22q11.2DS.

Performing pathway analysis on individual significantly differentially methylated probes can help provide an overview of the DNA methylation changes associated with the observed phenotypes, it is potentially more informative to focus on differentially methylated regions (DMRs) composed of multiple adjacent probes; potentially providing more confidence in the results. In this study 66 DMRs were detected, of which 48 were located in the promoter region of a gene and thus potentially having methylation levels directly correlated with its expression. Genes associated with these DMR are marked * below. Multiple DMRs are of particular interest here in the context of 22q11.2DS and its neuropsychiatric symptoms. *MIB2**, associated with the most significant DMR in this study, is involved in NOTCH signalling [355] which is associated with schizophrenia [356]. *NBEA** [357], *SORBS2** [357], *FILIP1** [358] and *TRAK1** [359] have been shown to have *de novo* mutations in schizophrenia patients. Deleterious variants in *TRAK1** also lead to lethal neurodevelopmental defects [360] and the protein product forms a complex with *DISC1* [361], a gene whose disruption has been shown to be associated with schizophrenia [362]. DISC1 regulates the GSK3 pathway and appears to control *AXIN2** levels [363]. Finally, *DPPA5** has also been found to contain a DMR in schizophrenia patients, but with the opposite direction of effect than that observed in 22q11.2DS patients [263]. No DMR, but two DMPs, are located within regions associated with schizophrenia in the largest GWAS study to date [92] (Figure 3-13).

NFIA and *NFIX* are the two most hypomethylated regions in the 22q11.2DS patients (Figure 3-12.D). They are both part of the transcription factor family of nuclear factor I (NFI) proteins, and mouse knock-out models of these genes present with abnormal brain development [364,365]. In humans, both genes are spanned by CNVs (1p31.3-p32.2 deletion for *NFIA* and 19p13.13 deletion for *NFIX*) that have been associated with brain malformations among other phenotypes [366,367]. Interestingly, 3 DMRs identified in this study are detected within the 19p13.13 deletion locus, overlapping the promoter regions of *NFIX*, *MAST1*, *AC02093.1*, *HOOK2* and *CTD6L21.1* genes. 8 probes are also individually significantly differentially methylated, and are associated with the genes *ZNF136*, *MAST1*, *DNASE2* and *NFIX*. It is however worth noting that no 19p13.13 deletion has been detected (methylation-array based detection) in any of the samples analysed in this study. Other DMR-associated genes that could be relevant to other 22q11.2DS phenotypes include *PLEKHB1* (ADHD candidate gene [368]), *HIC1* [369] and *HOXA2* [370] (developmental defects in mouse model, including palatal malformation for both genes), *DEAF1* (abnormal brain development and intellectual disability [371,372]), *SORBS2* (candidate gene within an intellectual disability causing CNV at locus 4q35.1 [373] and involved in dendritic development [374]), *RNF19A/DORFIN* (abnormal neurogenesis, memory and behaviour in mice knockout [375]) and *NBEA* (involved in synaptogenesis [376] and a candidate gene for autism [377]).

In this study, the DMPs within the A-D deleted region at 22q11.2 have been excluded due to the very strong differential methylation observed at the entire region. This potentially poses a problem of knowing if this window of association is due to a true biological signal or a technical artefact. To our knowledge, no such window of association around a hemizygous deletion has been previously reported, however it seems that only one other epigenome-wide epigenome case-control study of a CNV disorder has been reported. Analysis of the data from this study, looking at methylation changes in Williams Syndrome patients (7q11.23 deletion), revealed that a similar association is present at the hemizygously deleted locus. This association could result from a biological process either not known or not studied in this context. For example, biological mechanisms could compensate the reduced expression levels by decreasing DNA methylation levels at enhancer and promoter regions. That would imply a global hypomethylation of the hemizygously deleted region, but that does not seem to be the case at the 22q11.2 and 7q11.23 loci, where both hypo- and hyper-methylated probes were found. The deletion

of one chromosomal copy of these regions could have led to a difference in observed methylation levels if the regions were imprinted. Genomic imprinting is a mechanism that leads to genes being expressed differently on maternal and paternal chromosomes as a result of DNA methylation-mediated gene silencing [378]. The preferential deletion of the maternal (or paternal) copy of the 22q11.2 region could have led to such a bias in DNA methylation levels, depending on whether the silenced or expressed copy of the genes was present. While a number of studies have reported an increased maternal origin of the 22q11.2 deletion (~55%) [10,11], parent-of-origin specific expression studies of 22q11.2 orthologues in a mouse model (brain tissue) did not reveal any allelic specific expression of genes included within the smaller A-B deletion [379]. Moreover, no differences in parental origins of the deletion at 7q11.23 have been observed [10]. Together, these results and the findings from the current methylation study would suggest that the association observed at these deleted loci could be a technical artefact. It could potentially be caused by lower signal intensities of probes targeting CpGs at these hemizygously deleted regions, leading to an overestimation of the significance of small differences. Other DNA-methylation analysis techniques such as pyrosequencing or reduced representation bisulfite sequencing could potentially to explore this issue further and try to investigate the origin of the signal (biological or technical). It needs to be addressed because recent estimations have shown that up to 10% of the genome could contribute to copy number variants (CNVs) [155]. The presence of CNVs is commonly checked in genome-wide associations studies (GWAS), but mainly for CNVs known to affect the phenotype/disorder tested in the GWAS. This is less common in EWAS studies, due to the potential lack of genotyping information for the samples. The present study would imply that potentially any CNV, associated with the phenotype of interest or not, could have an impact on DNA methylation and potentially introduce a bias if its presence is unbalanced between cases and controls. For this reason, it is important to investigate this issue by performing CNV case-control epigenome wide association study (EWAS) of different CNVs and using different techniques of DNA methylation measurement.

The changes in DNA methylation levels observed in this study outside of 22q11.2 are likely due to the transcriptome changes caused by the 22q11.2 deletion. Changes in the expression of genes coding for proteins such DNMTs or transcription factors could impact on DNA methylation (cf. Figure 3-2) indirectly. However, these changes would not be specifically associated to schizophrenia. In order to explore this association a case

control EWAS comparing DNA methylation levels of 22q11.2DS patients with or without schizophrenia would need to be performed. Possible differences in DNA methylation between these two groups could be due to variants outside the 22q11.2 associated with schizophrenia and affecting DNA methylation directly or indirectly; they could also be due to gene-environment interaction altering DNA methylation. These interactions could be potentially causative in the increased the risk to schizophrenia. It has been shown that environmental factors such as stress or cannabis use could have an effect on schizophrenia symptoms and endophenotypes [380]. The gene-environment interactions leading to changes in DNA methylation could however also be a consequence of schizophrenia. For instance, Olanzapine, an atypical antipsychotic, leads to an alteration of DNA methylation levels of genes within the 22q11.2 deletion in different tissues [292]. For this reason, it would be important to assess any potential confounder, such as drug use/treatments and correct for it in such an EWAS.

Finally, despite the advantage of easy availability and the potential for disease biomarker discovery, the use of DNA extracted from blood has two major disadvantages in the context of 22q11.2DS. First, the deficit in T-cell observed in the 22q11.2DS patients led to an imbalance in cell-type proportion that is reflected at the DNA methylation levels. This introduced an important bias, and while it was corrected, this correction could have resulted in an increased rate of false-negative associations (due to overcorrection) while still retaining false-positive associations (imperfect correction, as the KEGG pathway enrichment in Hematopoietic cell lineage genes might suggest (Figure 3-11)). Secondly, the main phenotype of interest in this study was schizophrenia and as such, DNA extracted from brain tissue would have been more biologically relevant. While a correction for neuronal specific cell-type composition would still be needed, it is more likely that schizophrenia-associated DMPs and DMRs from brain tissue are linked to relevant gene pathways. However, despite these limitations it is encouraging that this study has identified DMPs that have been previously shown to have correlated methylation levels between blood and brain tissue [321].

The present study is the first EWAS of 22q11.2 deletion syndrome patients. The results show enrichments in pathways that are biologically relevant, in particular for schizophrenia, and differentially methylated regions in promoter regions of genes that could have an effect in this neurodevelopmental disorder. Replication studies with increased sample size, correlation with gene expression data and use of brain tissue will

help to provide a better insight into the biology of this syndrome. This will hopefully lead to the discovery of an epigenetic signature that might lead to better symptom prediction and/or treatments.

4. CRISPR/Cas9 genome editing of DGCR8 in human embryonic stem cells

4.1. Introduction

4.1.1. DGCR8, the microprocessor complex and microRNA processing

Among the genes deleted in the 1.5 mb-long 22q11.2 deletion – also called "DiGeorge syndrome chromosomal region (DGCR)" – *DGCR8* is of particular interest. Its hemizygous deletion may explain the wide range of symptoms observed in patients with 22q11.2 syndrome. First discovered in 2003 by Shiohama *et al.*, *DGCR8* has been shown to have a restricted expression pattern during development in tissues relevant to 22q11.2DS symptoms, such as the embryonic neuroepithelium, thymus and palate [381]. The DGCR8 protein was shown to interact with DROSHA, a nuclear RNAse (RNAse III protein). Together they form the microprocessor complex, an important component of the canonical microRNA biogenesis pathway [126] (Figure 4-1.A). MicroRNAs (miRNAs/miRs) are small non-coding RNAs (~22 nucleotides) involved in the regulation of gene expression. They are derived from primary microRNA (pri-miRNA) transcripts produced by the RNA polymerase II. These pri-miRNAs have a particular hairpin structure (also called RNA stem-loop) that is processed by the microprocessor complex into shorter precursor microRNAs (pre-miRNAs). These pre-miRNAs are transported out of the nucleus by the Exportin-5 (XPO5) protein. They are then trimmed by the endoribonuclease Dicer into double stranded miRNA/miRNA duplexes leading to the mature form of the miRNAs. miRNAs activate the RNA-induced silencing complex (RISC) in which the Argonaute II protein form a duplex between the miRNAs and a complementary mRNA. This leads to a regulation of gene expression through different repression mechanisms such as mRNA degradation or translational repression (as reviewed in Nilsen 2007 [382]). Some miRNAs are however also generated in microprocessor-independent pathways, for instance as a product of RNA splicing in which hairpin-containing introns (named mirtrons) generate pre-mRNAs [383].

DGCR8 protein is localised mainly in the nucleus – more particularly the nucleolus – but has also been found in the cytoplasm [384,385]. It contains two double-strand RNA binding domains (dsRBDs) that interacts with pri-miRNAs and are essential to the function of the microprocessor complex [386]. Its other domains include a nuclear localisation signal (NLS), a heme-binding / dimerization domain (WW / Rhed) and a

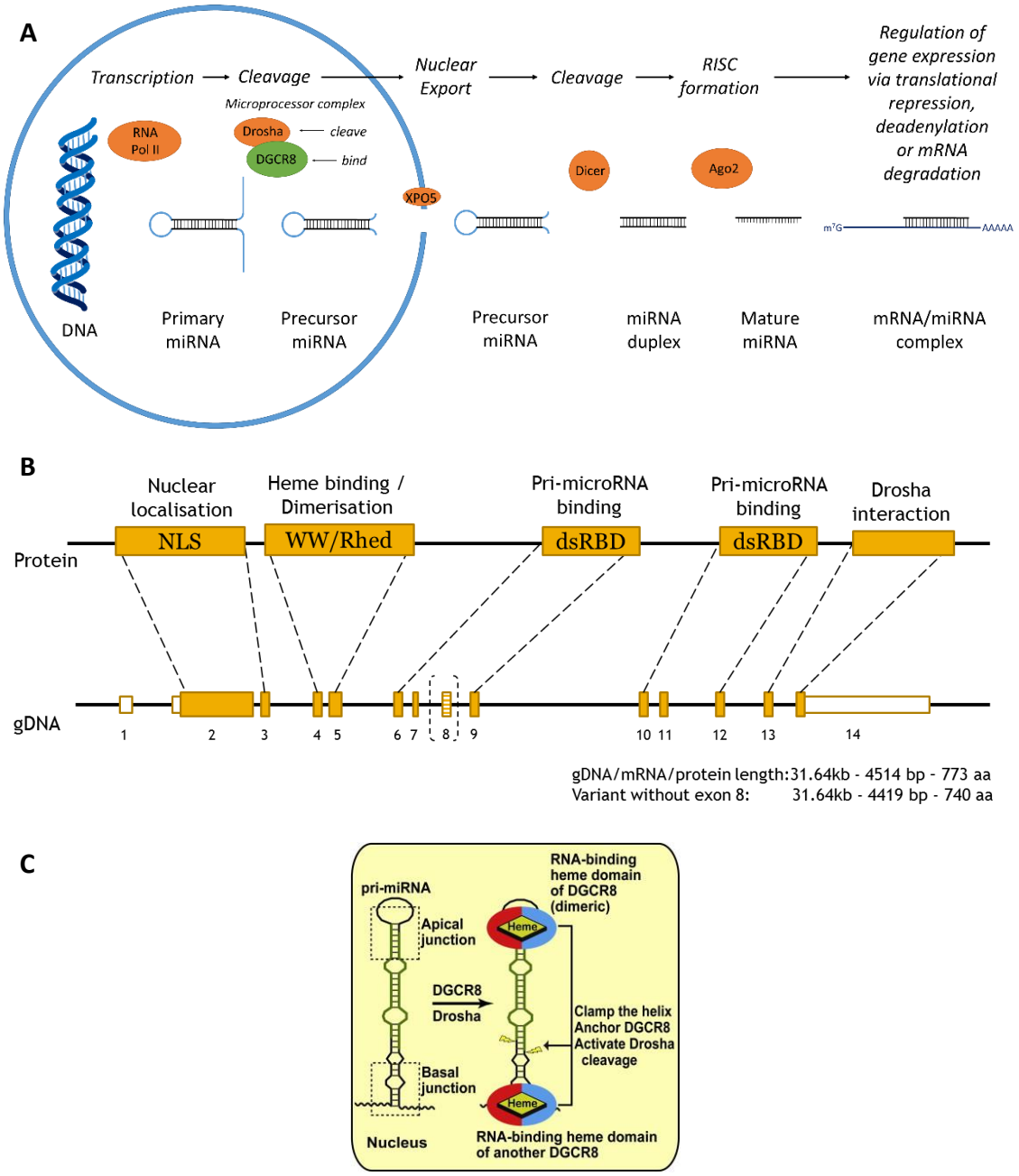


Figure 4-1. DGCR8 structure and activity. (**A**) **Canonical microRNA biogenesis pathway.** Proteins involved are represented in colored ovals. The blue circle represents the cell nucleus. (**B**) ***DGCR8* genomic DNA (gDNA) and protein structure.** Names of protein functional domains are indicated in boxes, with their function on top. Protein domains are depicted in rectangles mapped to the corresponding exons, with their function on top. Exons are depicted in full rectangles (coding exons) or empty rectangles (non-coding exons). dsRBD domain: double-stranded RNA-binding domain; gDNA: genomic DNA; NLS: nuclear localization signal; Rhed domain: RNA-binding heme domain. (**C**) **The DGCR8 RNA-Binding heme domain recognizes primary microRNAs by clamping the hairpin**. The DGCR8 subunits in a dimer are shown in red and cyan. The thick avocado strands represent 5′ and 3′ mature miRNAs. Figure (A) adapted from Brzustowicz 2012 [480], used under CC BY 3.0 license. Figure (C) from Quick-Cleveland 2014 [388], used under CC BY-NC-ND 3.0 licence.

domain on its C-terminal region that interacts with DROSHA [386,387] (Figure 4-1.B). The

Rhed domain (RNA-binding heme domain) is essential for forming a DGCR8 dimer that clamps to the apical or the basal junction of the primary microRNA to activate its cleavage by DROSHA [388] (Figure 4-1.C). It is worth noting that additional cofactors can bind to the microprocessor complex to regulate its activity as well as affecting which miRNAs are processed, as reviewed in Ha 2014 [389].

Despite being mostly known for its involvement in primary-microRNA processing, DGCR8 can also bind to other stem-loop containing RNAs that will be cleaved by DROSHA [390–393]; cross-linking immunoprecipitation experiments have shown that DGCR8 can bind to many classes of RNAs such as messenger RNAs (mRNAs) [390,392]. This extend further the capacity of DGCR8 to affect gene expression and as a consequence it potentiates the effects of gene dysregulation when *DGCR8* is hemizygously deleted such as in 22q11.2 deletion syndrome.

4.1.2. Animal models

The *DGCR8* gene appears to be relatively well-conserved and have orthologues in most species (Supplementary figure 4-1). Multiple animal models have been generated to study the impact of *DGCR8* deficiency, for instance in *C.elegans*, *D.melanogaster*, *D.rerio* and *M.musculus* as reviewed in Guna 2015 [74]. Homozygous knock-outs (KO) of *Dgcr8* is lethal before the end of pupal stage in *D.melanogaster* and during embryogenesis in *M.musculus* [74]; however the use of heterozygous and/or conditional KOs have allowed investigations into the impact of *Dgcr8* deletion in various aspect of development. The loss of microRNAs due to the deletion of *Dgcr8* has been shown to have an effect on tissues linked to 22q11.2 deletion syndrome symptoms, for instance DGCR8 KO in kidney epithelial cells leads to hydrophrenosis and kidney cysts [394] and to dilated cardiomyopathy when targeting cardiac myocytes [395].

Dgcr8 and *Drosha* conditional knock-out or knock-down in mice have previously shown that the microprocessor complex is essential for embryonic neurogenesis. It leads to precocious differentiation of neural progenitor cells (NPC) from the ventricular zone [396]. The microprocessor complex maintain the neural stem cell status and self-renewal by regulating pro-neuronal transcription factors' mRNA levels such as *Neurog2* [396] and *Tbr1* [397]. This effect seems to be mostly miRNA independent, as *Dicer* KO results in less impairment of corticogenesis than *DGCR8* KO [397]. Consistent with these results, Ouchi

et al. reported that *DGCR8* heterozygous KO mice exhibit reduced neural progenitor cell proliferation as well as cognitive and behavioural deficits [398]. These hippocampal NPCs proliferation and behavioural deficits can be rescued with Insulin-like growth factor 2 (IGF-2). Deletion of *Dgcr8* seems to also have an effect on mature neurons. Conditional deletion of *Dgcr8* in cortical pyramidal neurons leads to smaller neuronal size (resulting in loss in cortical thickness), deficits in inhibitory postsynaptic currents and a decreased abundance of parvalbumin interneurons in the prefrontal cortex [399]. *Dgcr8* heterozygous deficiency in mice results in impairment in short-term synaptic plasticity in the medial pre-frontal cortex (greater level of short-term depression) [400].

Importantly, *Dgcr8* knock-out mice seem to recapitulate some of the phenotypes observed in mouse models of the 22q11.2 deletion. Firstly, miRNA biogenesis seems to be altered in both *Dgcr8*$^{+/-}$ mice and *Df(16)A*$^{+/-}$ mice, a mouse model that mimic the 1.5 mb human 22q11.2 deletion (LCR A to LCR B) with the hemizygous deficiency of 27 genes [78]. Even though *Dgcr8*$^{+/-}$ mice did not show the exact same behavioural and cognitive deficits compared to the 22q11.2DS model, they both have deficits in the pre-pulse inhibition task which is considered to be an endophenotype of schizophrenia in human and mice [79]. They both also present altered spatial working memory–dependent learning. *Dgcr8*$^{+/-}$ mice and 22q11.2DS models *Df(16)1/+* and *Df(16)2/+* have an age-dependent long-term potentiation (LTP) increase which can be rescued in *Dgcr8*$^{+/-}$ by reintroducing microRNAs miR-25 and miR-185 that regulate *Serca2* mRNA levels (coding for a Ca^{2+} ATPase) [83]. Cortical interneurons and hippocampal dentate precursors have been shown to have a altered migration in both *Df1/+* and *Dgcr8*$^{+/-}$ mice, due to abnormal *Cxcr4*/*Cxcl12* signalling [401]. This phenotype can be partly rescued in *Df1/+* mice by reintroducing DGCR8 (lentiviral expression) in medial ganglionic eminence (MGE)-derived cells [401]. Finally, *Df(16)1/+* and *Dgcr8*$^{+/-}$ mice auditory thalamus have increased levels of D2 dopamine receptor *Drd2*, an antipsychotic molecular target, and their auditory thalamocortical projections are hypersensitive to antipsychotics [402]. Taken together, these evidences support the relevance of *DGCR8* haploinsufficiency as a model to investigate the aetiology of neuropsychiatric symptoms in 22q11.2 deletion syndrome.

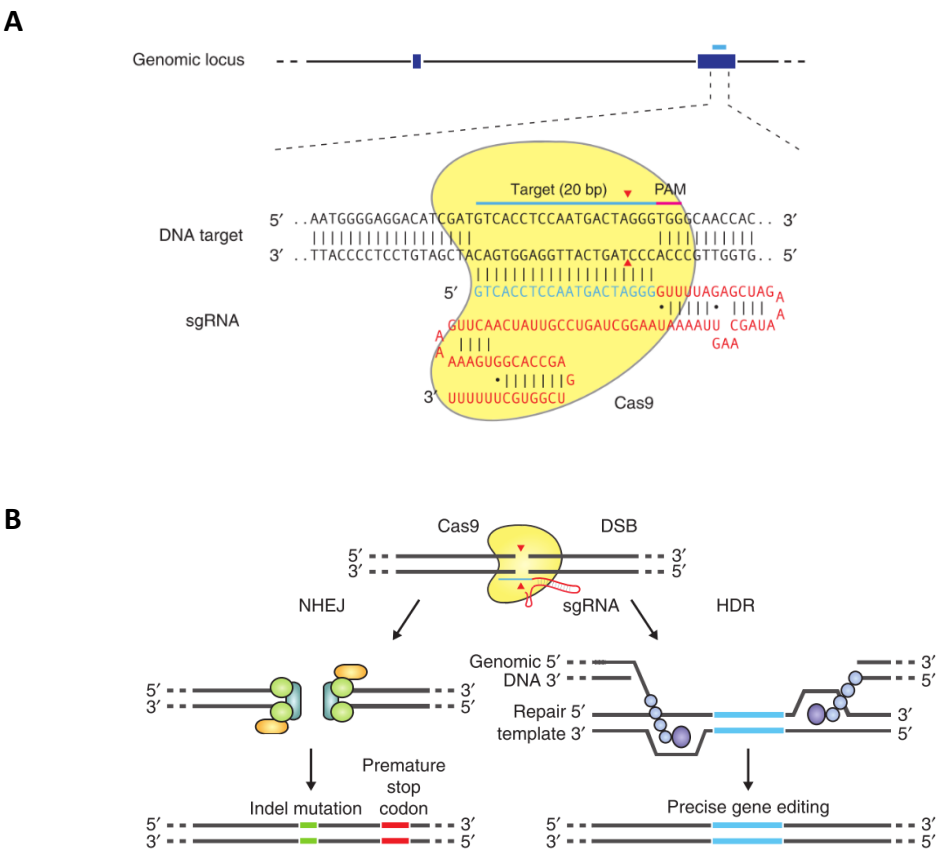


Figure 4-2. CRISPR/Cas9 system. (**A**) **Schematic of the RNA-guided Cas9 nuclease.** The Cas9 nuclease from S. pyogenes (in yellow) is targeted to genomic DNA (shown for example is the human EMX1 locus) by an sgRNA consisting of a 20-nt guide sequence (blue) and a scaffold (red). The guide sequence pairs with the DNA target (blue bar on top strand), directly upstream of a requisite 5′-NGG adjacent motif (PAM; pink). Cas9 mediates a DSB ~3 bp upstream of the PAM (red triangle). (**B**) **Double Strand Break (DSB) repair promotes gene editing**. DSBs induced by Cas9 (yellow) can be repaired in one of two ways. In the error-prone NHEJ pathway, the ends of a DSB are processed by endogenous DNA repair machinery and rejoined, which can result in random indel mutations at the site of junction. Indel mutations occurring within the coding region of a gene can result in frameshifts and the creation of a premature stop codon, resulting in gene knockout. Alternatively, a repair template in the form of a plasmid or ssODN can be supplied to leverage the HDR pathway, which allows high fidelity and precise editing. Single-stranded nicks to the DNA can also induce HDR. HDR: homology-directed repair; NHEJ: non-homologous end-joining; PAM: Protospacer adjacent motif; sgRNA: single-guide RNA; ssODN: single-stranded oligodeoxynucleotides. Figures and legends reprinted with permission from Springer Nature: Nature, Ran 2013 [406], copyright (2013).

4.1.3. Pluripotent stem cell models

Advances in genome editing techniques and in-vitro cell differentiation protocol have made disease modelling with human embryonic stem cells an attractive method to elucidate the underlying mechanisms leading to human disease [184]. *Dgcr8* knock-out mouse embryonic stem cell (mESC) have previously been generated and studied. Consistent with DGCR8's role in the microprocessor complex, *Dgcr8* KO mESCs have an altered microRNA biogenesis. A subset of precursor and mature microRNAs are

absent in *Dgcr8* $^{-/-}$ cells compared in *Dgcr8* $^{+/-}$ and wild-type cells; and more generally a global reduction of miRNA levels is observed in *Dgcr8* KO cells [128,403]. Moreover, *Dgcr8* KO hESCs have a deficit in proliferation characterised by an accumulation in the G1 phase of the cell cycle. They also exhibited an abnormal differentiation and do not silence entirely pluripotency markers such as *Oct4* or *Nanog*. Reintroducing embryonic stem cell specific miRNAs that are part of the miR-290 family can rescue the proliferation phenotype, possibly by interacting with genes of the cyclin E–Cdk2 pathway such as *Cdkn1a* or *Rlb2* [404]. More miRNAs have since been similarly implicated in embryonic stem cell self-renewal by studying *Dgcr8* KO mESCs; in particular miR-27a and miR-24 [405]. The dysregulation of gene expression in *Dgcr8* KO mESCs is discussed in in more details in Chapter 5 (p.106).

No DGCR8 KO in human embryonic stem cell have been published to date. The development of CRISPR/Cas9 genome editing system now permits to generate knock-out in human cells an efficient and cost-effective way. This system is derived from the bacterial *S.pyogenes* Cas9 nuclease and allows precise genome editing by inducing targeted double-strand DNA break directed by specifically designed guideRNAs [406,407] (Figure 4-2).

4.2. Aims of the chapter

In this chapter, I present the generation of *DGCR8* KO lines in human embryonic stem cells (hESCs) using the CRISPR/Cas9 technology. The goals of this project are:

- To obtain stable *DGCR8* KO hESCs.
- Determine how DGCR8-deficiency affects RNA processing ability.

This works will serve as the basis of future transcriptomic studies to investigate the extent of *DGCR8* contribution to the dysregulation of gene expression in 22q11.2 deletion syndrome.

4.3. Methods

4.3.1. Cell lines

The H7 (WA07) human embryonic stem cell (hESC) line has been used in this study (WiCell Research Institute, Inc.). This cell line was derived from a female subject with a

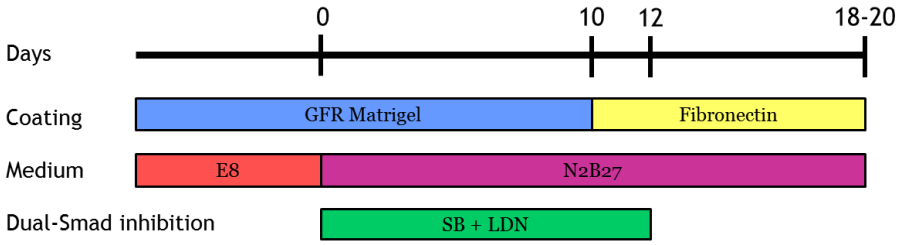


Figure 4-3. **Neural differentiation protocol.** Dual-SMAD signalling inhibition has been used to obtain neural progenitor cells, following a cortical neurodifferentiation protocol. Media and coating compositions are detailed in the methods.

karyotype of 46XX, European ethnicity (closest reference population: Tuscan/Palestine), B blood type, MHC class I antigens positive, MHC class II antigens negative). This line is also referred to as Wild-Type (WT) in this thesis.

4.3.2. hESC culture and neural differentiation

hESCs were cultured in Matrigel matrix-coated plates (Corning) and TeSR-E8 feeder-free medium (Stemcell technologies) changed every day and passaged with Gentle Cell Dissociation Reagent (Stemcell technologies) at ~80% confluency. Neural differentiation was induced using a modified version of the dual inhibition SMAD signalling protocols developed and described by Chambers *et al.* [185] (similar to the control condition in Arber *et al.* [408]). The differentiation medium used, N2B27, was composed of 1/3 Neurobasal medium, 2/3 DMEM/F-12, 1X N-2 supplement, 1X B-27 supplement, 20 mM L-glutamine, 20 µm β-mercaptoethanol (all from ThermoFisher Scientific) and 1X MycoZap Plus-CL antibiotics (Lonza). Neural differentiation media was changed every other day. Briefly, cells were plated in E8 medium in 12-wells plates coated with growth factor reduced (GFR) matrigel (Corning). After 24-48h, media was replaced with N2B27 medium supplemented with 10 µM SB-431542 (SB, TGF-B inhibitor) (StemCell technologies) and 100 nM LDN-193189 (LDN, BMP4 inhibitor) (Sigma-Aldrich) till day 12; cells were passaged to fibronectin-coated 12-wells plates (15µg/mL) at around day 9-10. Neural progenitor cells were collected at days 18 to 20. A summary of this differentiation protocol can be seen in Figure 4-3.

4.3.3. Genome editing

Editing of *DGCR8* gene has been performed using Zhang lab CRISPR/CAS9 protocol rev20140509, plasmid pX330 (pX330-U6-Chimeric_BB-CBh-hSpCas9 (Addgene plasmid #42230, deposited by Pr. Feng Zhang)) [406]. Two guide RNAs (gRNAs) were designed with an online CRISPR design tool (crispr.mit.edu, Zhang lab) (quality score > 90, guide sequences CACTACAGTTCGGGTCTATG (AGG) and CTCATAGACCCGAACTGTAG (TGG), confirmed with CRISPR/Cas9 target online predictor CCTop [409]). Two extra Guanine have been added at the 5' end to reduce off-target activity [410]. The guideRNAs (gRNAs) were then inserted into pX330 plasmids. gRNA efficiency was tested in HEK293 cells by transfection of the modified pX330 plasmids with lipofectamine 3000 (ThermoFisher Scientific)). The region targeted by the gRNAs were amplified by PCR, sequenced and subsequently analysed with the Tracking of Indels by Decomposition TIDE online tool (tide.nki.nl [411]) to assess gRNA efficiency (results not shown). Repair template for homology-directed repair (HDR) has been created using a pPGKpuro backbone (Addgene plasmid #11349, constructed by Peter W. Laird, deposited by Rudolf Jaenisch) to insert a puromycine resistance cassette at the targeted site and allow selection of recombinant clones. Briefly, two homology sequences of around 900 bp on both sides of the locus targeted by the gRNAs were amplified by PCR and inserted directionally with double digests/ligations. Details of pX330 and pPGKpuro plasmids and their editing can be seen in Figure 4-4.B/C. Restriction digests were performed using New England Biolabs restriction enzymes and reagents. Plasmid transformation was done in homemade chemically competent *E. coli* cells by heat shock following Invitrogen's One Shot TOP10 protocol. Double transfection (pX330 + pPGKpuro) was performed on H7 cells using a 4D-Nucleofector system (Lonza). Various plasmid ratios and total input DNA amount have been used. H7 cells (WT) were disrupted to single cells before nucleofection and then plated into 6-well plates. Cell survival was increased by using the RevitaCell supplement (ThermoFisher Scientific). After 24 to 48h, puromycine (Sigma-Aldrich) was added at 0.5 µg/ml then 0.1 µg/ml after 3 days. Cell colonies were subsequently picked manually and cultured in 48-well plates. PCR screening strategy is described Figure 4-4.D/E.; for clones with single-allelic insertion of the resistance cassette, the second allele was sequenced by GATC Biotech. PCR was performed using either HotStarTaq DNA polymerase (Qiagen) for cloning and sequencing or MyTaq DNA polymerase (Bioline) for PCR screening.

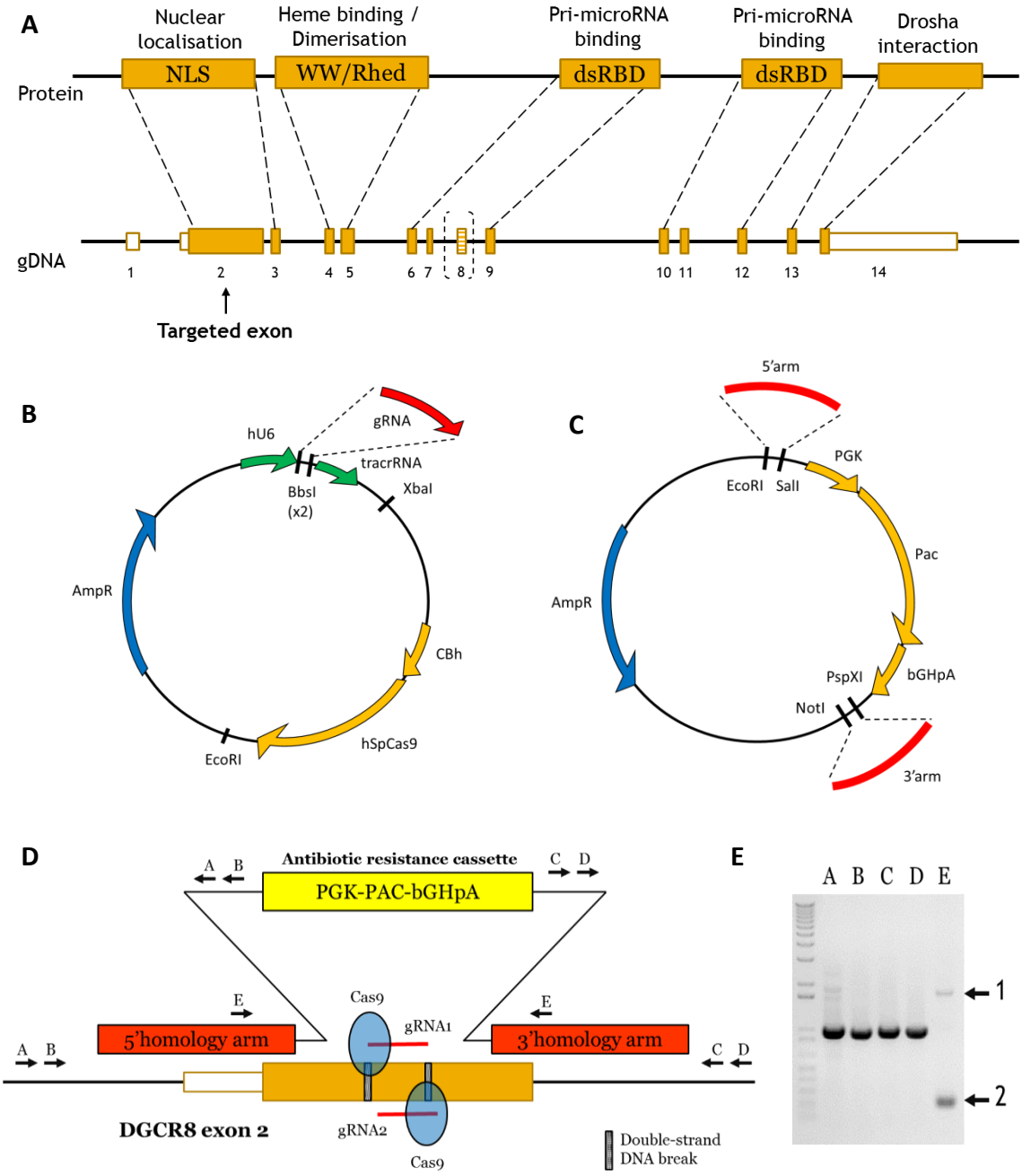


Figure 4-4. Genome editing strategy using CRISPR/CAS9 technology. Legend on next page.

4.3.4. RNA extraction and qRT-PCR

Cells were lysed with TRI reagent (Sigma-Aldrich); then RNA was extracted by addition of 0.2 vol. chloroform, precipitation with 0.7 vol. isopropanol followed by two 70% ethanol washes/precipitation. RNA pellets were air-dried then resuspended in

Figure 4-4. Genome editing strategy using CRISPR/CAS9 technology. (**A**) DGCR8 genomic DNA (gDNA) and protein structure. **(B**) Plasmid pX330 (Addgene, Zhang lab) containing Cas9 sequence and insertion of guideRNA at the BbsI restriction sites. (**C)** Plasmid PGKPurobpA and insertion of 5' and 3' homology arms for homology-directed repair (HDR) on exon 2 of DGCR8 (873bp and 892bp respectively). (**D**) Genome editing and PCR screening strategy. The nuclease Cas9 induces a double-strand break at the target site recognised by the gRNA that will be repaired by HDR with the targeting vector, allowing insertion of a selection cassette (puromycine resistance, PAC gene). Arrows labelled A to E represent the primers set used for PCR screening. Correct cassette integration was detected using forward primer A and reverse primer B, then confirmed with primer sets A to D. Primer set E was used to determine zygosity status. (**E)** Example of PCR screening results for a heterozygous recombinant clone. Expected sizes for primers sets A to D: 996bp, 971bp, 992bp, 997bp respectively; primer set E: 1748bp with cassette insertion (band 1), 248 bp without (band 2). The band 2 was then sequenced to determine possible insertion/deletion mutations due to non-homologous end joining repair errors. *Abbreviations: 5'/3' arms: homology arms for HDR at DGCR8 exon2 locus AmpR: ampicillin resistance. BbsI, EcoRI, NotI, PspXI, SalI: restriction enzyme sites. bGHpA: Bovine Growth Hormone Polyadenylation Signal (termination signal). CbH: CBA promoter. dsRBD: double-stranded RNA-binding domain.gRNA: guide RNA. hSpCas9: humanized S.pyogenes Cas9. hU6: humanized U6 promoter. NLS: nuclear localization sequence. Pac: Puromycin resistance gene. PGK: PGK promoter. tracrRNA: transactivating RNA. WW: WW domain.*

DEPC-treated water and quantified/qualified with an Eppendorf BioSpectrometer. Samples with a 260/280 ratio of 1.8-2.2 were kept for subsequent analyses. Some samples used in Figure 4-9 (A) and (B) had 260/280 ratio < 1.8; visual inspection of qPCR amplification and melting curves was performed for these samples. RNA was treated by Turbo DNAse digestion (Ambion, Life Technologies) (Figure 4-6, Figure 4-7 and Figure 4-8) or by PerfeCTa DNAse I (Quantabio) (Figure 4-9) then retrotranscription was performed with the qScript cDNA Synthesis Kit (Quantabio) (Figure 4-6, Figure 4-7 and Figure 4-8) or qScript cDNA SuperMix kit (Quantabio) (Figure 4-9). Samples were diluted 1:100 final in nuclease-free water and 1X MESA GREEN qPCR MasterMix Plus for SYBR Assay (Applied Biosystem, Life Technologies) with 1 µM primers (primer list in Supplementary table 4-1). Quantitative real-time PCR was executed on a CFX Connect Real-Time PCR Detection System (Bio-Rad) and analysed with the CFX Manager Software (Bio-Rad). PCR primers were designed with Primer3Plus (unless specified otherwise in Supplementary table 4-1) and produced by Sigma-Aldrich. The number of technical and biological replicates used are detailed in each figure. Calculations of relative expression were performed using the $2^{-\Delta\Delta CT}$ method [412] using a custom R-script (Microsoft R Open 3.4.0 (Microsoft), based on R-3.4.0 (R Statistics)). Statistical analysis was performed on ΔCt values with the Mann-Whitney-Wilcoxon Test (*wilcox.tes*t from R package *Stats*). qPCR plots were generated with the R package *ggplot2* [242].

4.3.5. Immunostainings

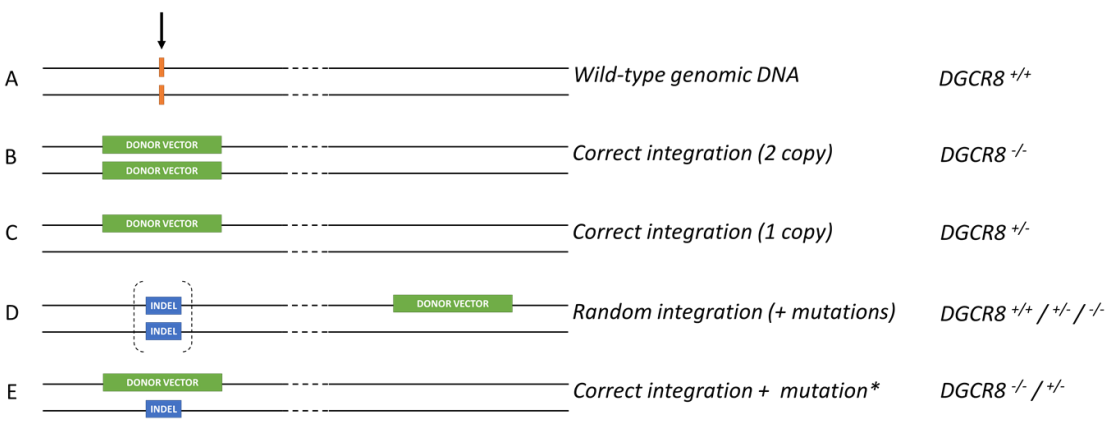
Cells were washed with PBS twice and fixed with 3.7% Paraformaldehyde for 20 minutes at room temperature. Cells were then washed twice (5 min each) with PBST (PBS + 0.3% Triton X-100 (ThermoFisher Scientific)) and blocked with 2% BSA (Bovin serum albumin, Sigma-Aldrich) and 5% donkey serum (Gentaur) in PBST (20 min). Primary antibodies were incubated in PBST + 5% donkey serum at 4°C overnight, cells were then washed 3 times for 20 min in PBST at room temperature before being incubated with secondary antibodies for 2h at room temperature in the dark. After a PBST wash (10 min) cells were counterstained with DAPI (1:3000 in PBST) (ThermoFisher Scientifc)) for 5 min. After 3 PBST wash (10 min each), all liquid was removed from the wells and cells were mounted with coverslips using Dako fluorescence mounting medium (Agilent). Primary antibodies and dilutions used were: OCT-3/4 goat Ab (Santa Cruz Biotechnologysc-86289, 1/500), TRA1-81 mouse Ab (Merck Millipore Chemicon MAB4381, 1/500), N-cadherin (NCAD) mouse Ab (Invitrogen 18-0224, 1/1000). Secondary antibodies were used at the dilution 1/500 (Alexa Fluor, ThermoFisher Scientific).

4.4. Results

4.4.1. Generation of DGCR8 knock-out human embryonic stem cells lines

The CRISPR/Cas9 genome editing technology was used to knock-out *DGCR8* in human embryonic stem cells. A custom donor plasmid was created to insert an antibiotic (puromycine) resistance gene in *DGCR8* gene by directed homologous recombination. The goal was to disrupt *DGCR8* and to be able to perform a positive selection on edited cells. *DGCR8* Exon 2 was targeted, which is DGCR8 first coding exon (Figure 4-4.A/D). Two guideRNAs targeting the same region (on opposite strands) were designed to increase the chance of DNA cleavage. Nucleofection strategy was optimised by changing total quantity and ratio of donor vector and gRNA/Cas9 containing plasmids (gRNA1 plasmid and/or gRNA2 plasmid) (data not shown). Use of linearized donor vector sequence (Cut by *EcoRI* and *NotI*, cf. Figure 4-4.C) led to more than 130 antibiotic-resistant clones with integrated cassette; however PCR screening (cf. Figure 4-5) revealed around 95% of random integration of the cassette in other genomic loci. For this reason,

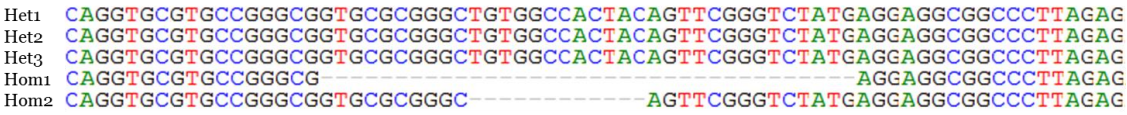
A



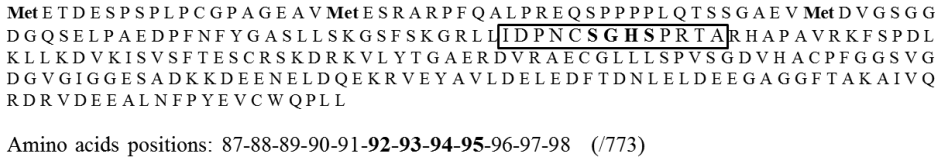
B

Donor vector	Linearized	Circular
Antibiotic resistant clones per transfection (~1-1.5* 10^6 cells per transfection)	15-30	1-3
Clones screened (B+C+D+E)	134	40
Clones with correct integration of donor vector (2 copy) (B)	0	0
Clones with correct integration of donor vector (1 copy) (C)	7	8
Clones with donor vector + NHEJ mutations (E)	0	2
Clones with donor vector + NHEJ mutations leading to a premature stop codon (E)	0	0

C



D



Amino acids positions: 87-88-89-90-91-**92-93-94-95**-96-97-98 (/773)

Figure 4-5. Results of genome editing and screening. (**A**) **Schematic of possible genome editing results and resulting genotypes**. INDEL: Insertion/deletion mutation (**B**) **Results and comparison between uses of linearized or circular donor vector**. (**C**) **Sequencing results** (from PCR amplicon with primer set E presented in Figure 4-4.E lane 2) of wild-type target sequence, example of one heterozygous knock-out and the two cell lines with INDEL mutations (Hom1 and Hom2 with 36bp and 12bp deletion respectively). (**D**) **Amino-acid sequence of DGCR8 exon 2**. In the box are presented the deleted amino-acids in Hom1, in bold those in Hom2. NHEJ: non-homologous end-joining.

the circular form of the donor plasmid was used. This strategy led to a greater specificity of integration (around 20%, 40 clones tested) (Figure 4-5.B).

Clones with the correct integration in one allele (cf. Figure 4-5.A potential genotype D-E) were sequenced to detect potential insertion/deletion mutation (INDEL) generated

via non-homologous end-joining repair in the other allele. The sequencing results for lines that are subsequently used in this thesis are presented in Figure 4-5.C. Two clones, named clone Hom1 and clone Hom2, had deletion mutations in the 2nd allele, which were 36bp and 12bp deletions respectively. However, no frameshift was generated by these mutations (as determined by ExPASy Translate Tool) (Figure 4-5.D). PROVEAN (Protein Variation Effect Analyzer [413]) predicts a neutral effect of the mutation in clone Hom2 (score = -1.28) but deleterious in clone Hom1 (score = -7.57). Moreover, SNP id rs182736423 (Chr22:20086250, C/T), contained in amino-acid 96 (deleted in clone Hom1) is characterised on ENSEMBL.org as "deleterious – low confidence" (SIFT prediction) and "probably damaging" (Polyphen prediction). This could suggest that the allele containing this 36bp deletion potentially code for a protein with an altered function. For this reason, this clone was preferentially used (over the clone Hom2) in the following experiments; clone Hom1 is also referred to as *DGCR8* $^{MUT/-}$ in this thesis.

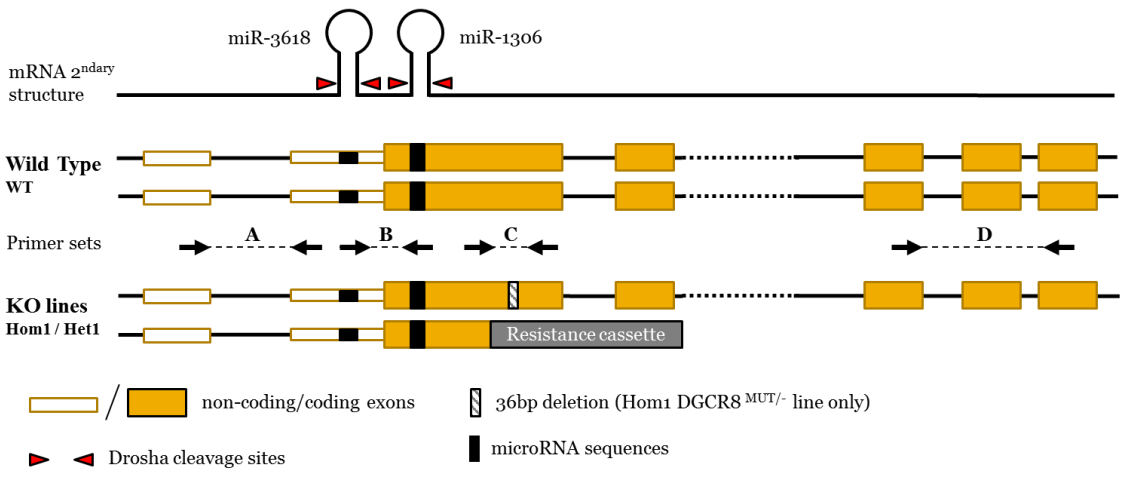
4.4.2. Validation of the model: preliminary study of expected results

4.4.2.1. Posttranscriptional regulation of DGCR8 mRNA

Recent studies have shown that DGCR8/DROSHA can bind and process RNAs that are not microRNAs, including hundreds of mRNAs [390–393]. Interestingly, it has been shown in *Drosophila* and mouse models that *Dgcr8* itself is among these mRNAs; there is a posttranscriptional control of *Dgcr8* expression by the microprocessor complex [391,414,415]. The exon 2 of *Dgcr8* contains two hairpins structures that are cleaved by the protein DROSHA to generate two microRNAs (miR-3618 and miR-1306), leading to a cross-regulation between DROSHA and DGCR8 (as DGCR8 protein stabilise DROSHA protein [415]). Downregulation of *DROSHA*, and thus of the microprocessor complex, lead to increased levels of *DGCR8* mRNA [415]. This feedback loop is reflected at the protein levels; brain tissue from *Dgcr8* heterozygous knock-out mouse have more than the expected 50% expression compared to controls (relative expression between 60% and 90% depending on brain regions and developmental time points) [78,416].

Different sets of primers (A, B, C, D) have been designed (as depicted Figure 4-6.A) to investigate the effect of the generated *DGCR8* mutations on this feedback loop by measuring RNA level at different *DGCR8* gene sites. The primer set A targets exons 1 to 2; primer set B spans the hairpin structure targeted by DGCR8; primer set C amplify the

A



B

Hypothetical situations leading to different mRNA levels		Primers sets			
		A	B	C	D
Posttranscriptional processing by DGCR8	ABSENT in KO lines	Hom1 = Het1 = WT	Hom1 = Het1 = WT	Hom1 = Het1 = ½ WT	Hom1 = Het1 = ½ WT
	PRESENT and ALTERED in KO lines	Hom1 ≥ Het1 > WT	Hom1 ≥ Het1 > WT	Hom1 ≥ Het1 > ½ WT	Hom1 ≥ Het1 > ½ WT

C

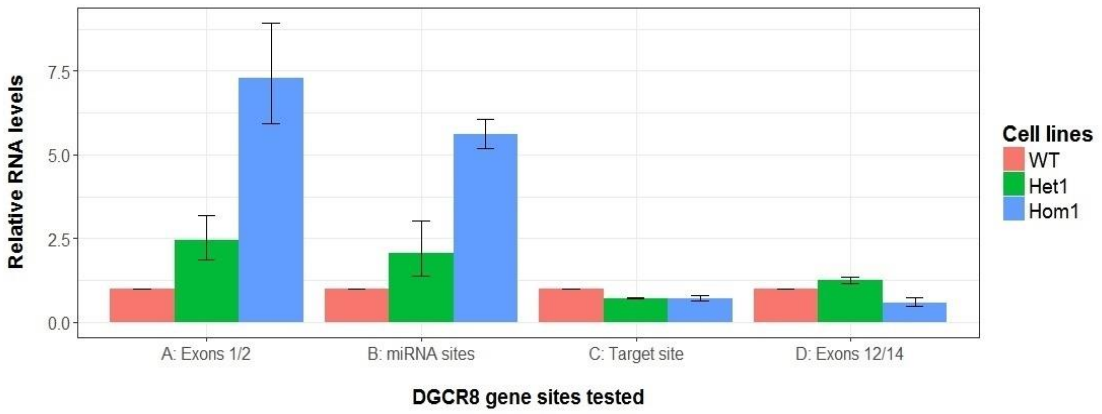


Figure 4-6. Alteration of *DGCR8* mRNA processing in embryonic stem cells. (**A**) Representation of precursor mRNA (before posttranscriptional processing), primer sets used to characterise mRNA levels and expected results. From top to bottom: mRNA secondary structure showing the hairpins contained in *DGCR8* exon 2, leading to excision by the microprocessor of 2 precursor microRNA; mRNA copies in Wild-type cells; Primer sets used for qPCR; mRNA copies in KO lines. (**B**) Table showing expected mRNA levels for the different primer sets depending on DGCR8 processing capacity in KO lines. (**C**) *DGCR8* mRNA levels as determined by qPCR using the primers sets described in (A). Normalised to GAPDH levels. N = 3 (technical replicates) x 2 (biological replicates) for KO lines, 3 (technical replicates) x 1 (biological replicates) for WT. Bars represent standard deviation between biological replicates. No statistical test performed due to sample size.

region edited by CRISPR; primer set D targets exons 12 to 14. Three cell lines were

compared: Wild-Type (WT), Het1 (*DGCR8* $^{+/-}$) and Hom1 (*DGCR8* $^{MUT/-}$).

Figure 4-6.B indicate the hypothetical relative mRNA levels between cell lines. It depends on the posttranscriptional ability of DGCR8 to process its own mRNA, which can be affected by the generated mutations. Primer sets A and B amplicons showed increased mRNA levels of around ~2- to ~2.5-fold between the heterozygous clone (Het1) and the control line (WT). This increase was greater for the lines Hom1 DGCR8 $^{MUT/-}$ (with a 36 bp deletion on second allele), with on average around ~5.5- to ~7.5-fold more mRNA detected compared to the control line. Primer sets C and D amplified regions on the other hand seemed to show less changes in mRNA levels. For the Het1 line, they indicated around ~30% average decrease and ~20% average increase respectively relative to WT. For the Hom1 line, there was a ~30% decrease for primer set C amplicon and a ~40% decrease for the primer set D amplicon.

The regions amplified by primer sets A and B were located upstream of the mutated sites; this increase possibly indicates an altered processing of *DGCR8* mRNA by the microprocessor complex (Figure 4-6.B). The mRNA levels observed could be due to reduced *DGCR8* mRNA degradation. The higher levels detected for the Hom1 line may indicate a further disruption of DGCR8-mediated processing. For the primer sets C and D amplicons only the intact allele is measured. In consequence, theoretically the mRNA changes compared to WT should be half of those measured for primer A and B amplicons. However observed levels are not concordant with this hypothesis. Alternative splicing of *DGCR8* or qPCR variability due to small sample size could explain the results.

4.4.2.2. microRNA processing

The main expected phenotype upon single copy deletion of *DGCR8* is an alteration of microRNA processing, as it has been shown in *DGCR8* knock-out models in *Drosophila* [417], mouse embryonic stem cells [128] and mice [78] as well as in 22q11.2 deletion syndrome mouse models and patients [129,418,419].

To assess whether the *DGCR8* KO lines had abnormal microRNA processing, a set of primary-microRNAs that were previously suggested to be processed by the microprocessor complex have been selected. These microRNAs include pri-miR 16-1, pri-let 7-a-1, pri-miR 185, pri-miR 20-19a and pri-miR 17. Pri-miR 16-1 and pri-let 7-a-1 have increased levels after *DROSHA* depletion in HEK cells [415]. Pri-miR 185 have been

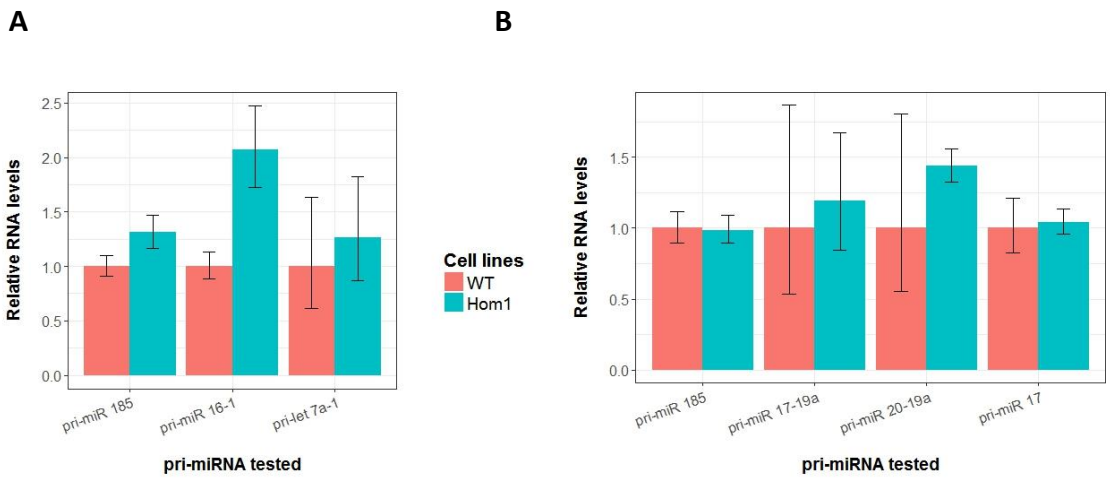


Figure 4-7. Precursors levels of a subset of microRNAs. (**A**) RNA levels in human embryonic stem cells. N = 3 (technical replicates) x 3 (biological replicates). (**B**) RNA levels in Neural Progenitor Cells (Day 20). N = 3 (technical replicates) x 2 (biological replicates). Normalised to GAPDH levels. Bars represent standard deviation between biological replicates. No statistical test performed due to sample size.

selected due to the decreased levels of miR 185 in *Dgcr8* KO mouse models [78]. Pri-miR 20-19a and pri-miR 17 are part of the miR 17~92 cluster that is affected by *DGCR8* deletion in vascular smooth muscle cells [420]. Moreover, microRNAs from this cluster have been shown to be implicated in neuronal differentiation, neural progenitor cell proliferation and neurogenesis [421–423].

These different primary microRNA have been tested in either undifferentiated hESCs (Figure 4-7.A) or neural progenitor cells at day 20 of differentiation (NPCs) (Figure 4-7.B). I only tested Hom1 KO line against the control line due to the predicted effect of the mutation on DGCR8 function and the greater changes in DGCR8 mRNA levels (cf. Figure 4-6.C). The greatest effect observed was for pri-miR 16-1, with a ~2-fold increase in undifferentiated hESCs. Other tested pri-miRNAs showed minor differences or none.

4.4.2.3. Neurogenin2 mRNA processing by the microprocessor complex

As discussed previously, the microprocessor complex has been previously shown to process a subset of mRNAs. Among these, an interesting transcript in the context of neural differentiation is coded by Neurogenin 2 (*NEUROG2*). It has been shown that DROSHA can cleave hairpin structures in the 3'UTR of *NEUROG2* mRNA and regulate its level [424,425]. NEUROG2 is a pro-neural transcription factor upstream of *NEUROD1* (also regulated by the microprocessor complex). It has been shown that enforced

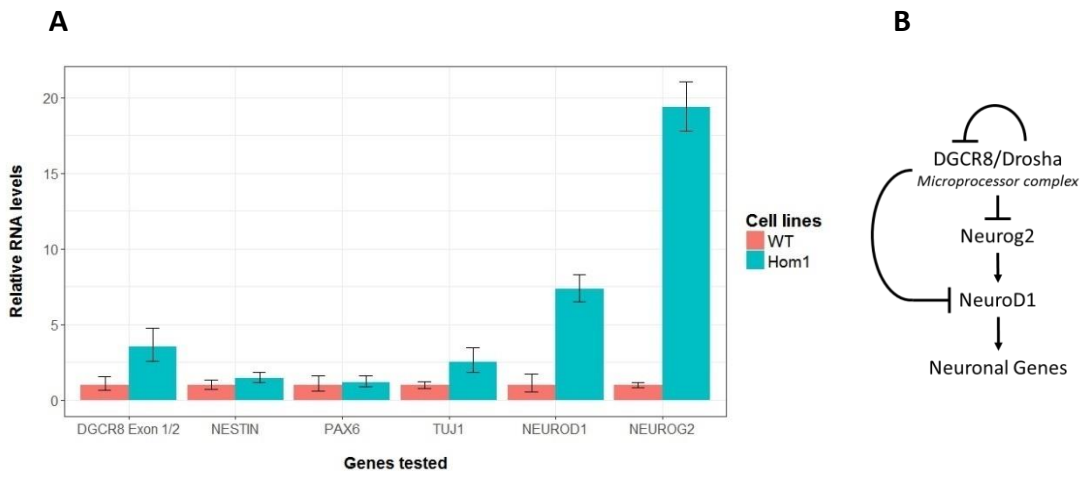


Figure 4-8. Microprocessor-dependent control of pro-neural mRNAs. (**A**). RNA levels in Neural Progenitor Cells (Day 20). N = 3 (technical replicates) x 2 (biological replicates). Normalised to *GAPDH* levels. Bars represent standard deviation between biological replicates. No statistical test performed due to sample size. *Marker used: Nestin: neural stem/progenitor cell marker; Pax-6: radial glia marker; TUJ1: Neuron-specific Class III β-tubulin (neuronal marker); NEUROD1, NEUROG2 (Neurogenin2): pro-neuronal transcription factors*. (**B**) Involvement of the microprocessor complex in a regulatory cascade controlling neuronal differentiation (adapted from Di Carlo *et al.* [425]).

expression of *NEUROG2* in human and mouse embryonic stem cells induce the cells to differentiate into neurons [426,427]. Furthermore, in human iPSC-derived neural progenitor cells, *NEUROG2* induction causes a rapid maturation of the cells [428]. Concordantly, inhibition of the microprocessor complex (*Drosha* inhibition) in forebrain neural progenitors lead to induction of *NEUROG2*, *NEUROD1* and result in early differentiation [424,425].

Using the primer set A (cf. Figure 4-6.A), I firstly examined DGCR8 mRNA level corresponding to exons 1 to 2. It showed an ~3.5-fold increase in day 20 neural progenitor cells (NPCs) DGCR8 KO lines compared to wild-type NPCs, similarly to what has been found at the embryonic stem cell stage (Figure 4-6). I then determined mRNA levels of *NEUROD1*, *NEUROG2* and different neural lineage markers in the Hom1 *DGCR8* $^{MUT/-}$ line. I observed an increase of both *NEUROD1* (~7.5-fold) and *NEUROG2* (~19.5-fold) in NPCs (day 20), as well as *TUJ*-1 mRNA level (neuron-specific marker, ~2.5-fold). However, the transcript level of *NESTIN* and *PAX6* (NPC and radial glia markers) were only slightly altered (~1.5- and ~1.2-fold increase respectively).

4.4.3. Results reproducibility

In order to assess the validity and reproducibility of the results, replications of the previous experiments have been done with a larger sample size (including different KO lines) and over multiple neural differentiation. Biological variation between independent neural differentiation and multiple cell lines is likely to introduce noise in the results; but the rationale of this experiment is to try and gain confidence on possible molecular phenotypes in the DGCR8 knock-out lines before pursuing with further experiments. While analysing results from single differentiation experiments independently could give a better picture of gene expression changes (such as what is shown in the preliminary results above), this will test the robustness of the results.

For this, 7 different cell lines have been used. This include the homozygous KO Hom1 (*DGCR8* $^{MUT/-}$), three distinct heterozygous knock-out lines Het1, Het2, Het3 (*DGCR8* $^{+/-}$) and three control lines (*DGCR8* $^{+/+}$). The first control line is the parental H7 line (WT) as used previously. A second line, named WTcrispr1, is a *DGCR8* wildtype sister clone of the above *DGCR8* KO lines derived from the same CRISPR/Cas9 genome editing process. This line has integrated the antibiotic resistance cassette but at a random and non-determined locus (cf. Figure 4-5.A hypothesis D). A third line, named WTcrispr2, is another *DGCR8* wildtype line of H7 background, derived from an independent CRISPR/Cas9 targeting study (provided by Dr. Lucia Cardo, Cardiff University). 3 biological replicates have been collected for each cell lines, and results from samples from the same conditions (WT, Heterozygous KO, Homozygous KO) were pooled as replicates and analysed together.

For the embryonic stem cell stage, 9 WT samples (3*3), 9 Heterozygous KO samples (3*3) and 3 Homozygous KO samples (1*3) have been collected. For the neural progenitor cell stage, three independent differentiations have been performed, resulting in 24 WT samples (9*3 minus 3 samples excluded due to abnormal melting curves/low cDNA levels), 27 Heterozygous KO samples (9*3) and 9 Homozygous KO samples (3*3). Due to other experimental constraints, NPCs were extracted at day 18 of differentiation, while samples for the preliminary qPCRs presented above were extracted at day 20.

At the embryonic stem cell stage, the differences in *DGCR8* mRNA levels previously observed were replicated (Figure 4-9.A), with increased RNA levels for the two primer

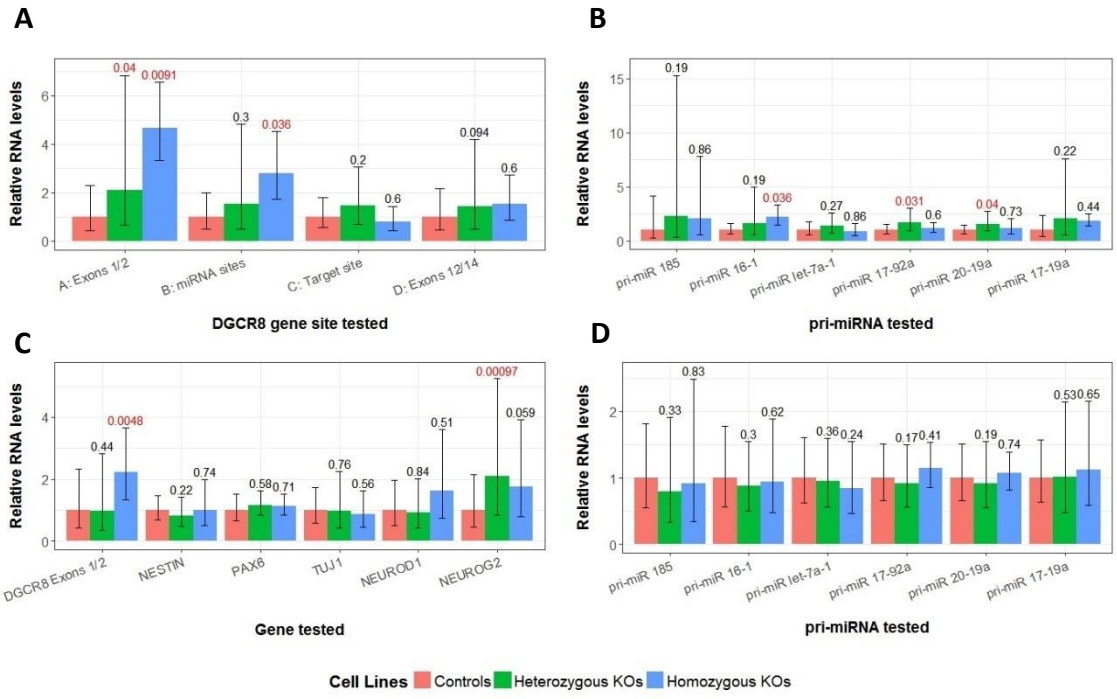


Figure 4-9. Results reproducibility. Quantitative PCR relative RNA levels for three controls lines (WT, WTcrispr1, WTcrispr2), three heterozygous *DGCR8* knock-out lines (Het1, Het2, Het3) and one potential homozygous *DGCR8* knock-out line (Hom1). For samples at day 0 (embryonic stem cells, (A) and (B)) N = 9 biological replicates for Controls and Homozygous KO lines, N= 3 biological replicates for the Homozygous line (no technical replicates for any line). For samples at day 18 (neural progenitor cells cells, (C) and (D)) N = 24 biological replicates for Controls, N = 27 biological replicates for Homozygous KO lines, N= 9 biological replicates for the Homozygous line (no technical replicates for any line; samples from 3 independent differentiations). Normalised to *GAPDH* levels (2 technical replicates for each biological replicates). Statistical testing performed on ΔCt values with the Mann-Whitney-Wilcoxon Test, without correction for multiple comparison. Bars represent standard deviation between biological replicates. **(A) Alteration of *DGCR8* mRNA processing in embryonic stem cells (day 0)** (cf. Figure 4-6). **(B / D) Precursors levels of a subset of microRNAs at (B) day 0 (embryonic stem cells) and (C) day 18 (neural progenitor cells)** (cf. Figure 4-7). **(D) Microprocessor-dependent control of pro-neural mRNAs at day 18** (cf. Figure 4-8).

sets (A and B) located before the edited region but no significant changes for the primer sets situated after the mutations (C and D). These differences were significant for both heterozygous and homozygous KO for the primer set A that spans *DGCR8* exons 1 to 2 (~2-fold increase, p=0.04 and ~4.5-fold, p=0.0091 respectively) respectively. However, it was only significant for the Hom1 KO for the primer set B spanning the hairpins sites processed by the microprocessor complex (~2.8-fold increase, p = 0.036). This difference in *DGCR8* mRNA level at the exons 1/2 locus was also significant in neural progenitor cells for the Hom1 clone (~2.2-fold increase, p = 0.0048) but not for the heterozygous KO lines (Figure 4-9.C).

Three primary microRNAs had significantly higher RNA levels in undifferentiated *DGCR8* KO lines (Figure 4-9.B): pri-miR 16-1 (~2.2-fold, p=0.036, Homozygous KO line) as well as pri-miRs 17-92a and 17-19a (~1.7-fold, p=0.031 and ~1.6-fold, p=0.04 respectively, Heterozygous KO lines). Two other pri-miRNA, pri-miR 185 and 17-19a, seemed to have increased RNA levels in the KO lines (~2-fold and 1.9-fold respectively) but the difference was not significant. The direction of effect for these different pri-miRNAs (increased levels) is consistent with a hypothetic disruption of DGCR8 function. At the neural progenitor cell stage however, no pri-miRNA appeared to be significantly impacted by the mutations in the different knock-out lines (Figure 4-9.D).

In NPCs, the mRNA levels of *NEUROG2* appeared to be increased in both heterozygous and homozygous lines. This difference was only significant in heterozygous KO lines (~2-fold, p=0.00097 and ~1.8-fold, p = 0.059 respectively) (*Figure 4-9.C).* None of the other transcript tested, *NEUROD1*, *NESTIN*, *PAX6* and *TUJ1*, appeared to have different mRNA levels in the KO lines.

4.4.4. Phenotypic observations

I did not observe gross changes in *DGCR8* knock-out KO cells at the embryonic stem cell stage (day 0). The KO cells did not appear to have increased level of spontaneous differentiation, or to exhibit deficit in cell viability or proliferation. The pluripotent cell markers OCT4 and TRA-1-81 were expressed by these cells in comparable proportion to that of the controls (cf. Figure 4-10).

However, at the neural progenitor cell stage (day 18) the *DGCR8* KO cells formed abnormal structures. At this stage of the differentiation there is presence of typical neural rosettes, which are 2-dimensional structures that are analogous to the neural tube that forms during development (cf. Figure 4-11.A). N-cadherin is commonly used to detect *in vitro* these round-shaped structures, as it marks their apical side. DGCR8 KO lines appeared to have relatively large and abnormal structures marked by N-cadherin (NCAD) (representative images shown Figure 4-11.B). Generally, differentiating *DGCR8* KO cultures appeared more heterogeneous than the control cultures with spaces without NPCs. Moreover, the abnormal NCAD+ structures tended to appear on the border of these spaces. The effect seemed to be greater in the homozygous knock-out Hom1; and it has been observed for multiple heterozygous knock-out in multiple differentiation runs.

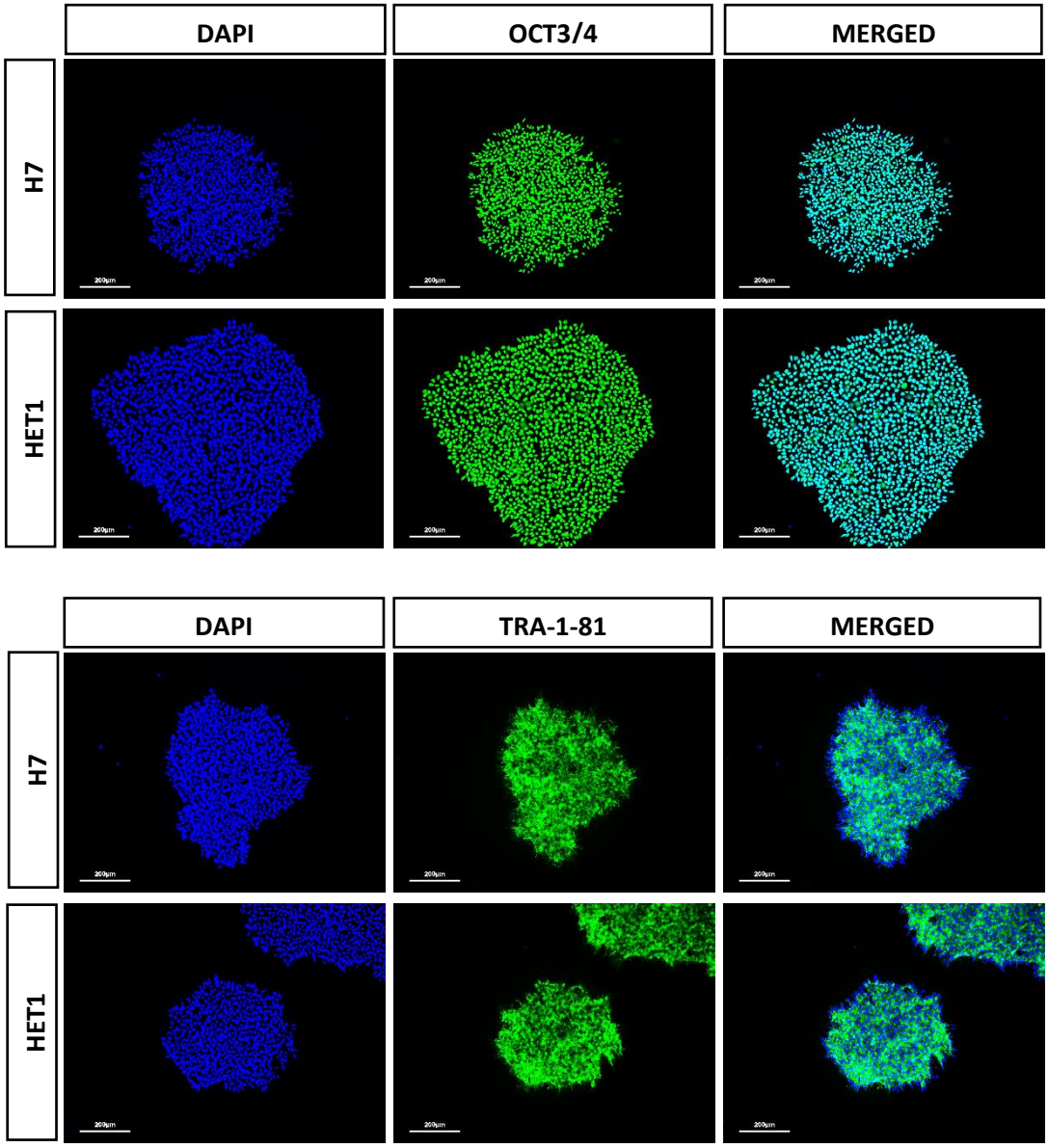


Figure 4-10. Pluripotency markers expressed at the stem cell stage. DAPI: nuclear counterstain; OCT3/4: nuclear marker of pluripotency; TRA-1-81: surface marker of pluripotency. WT: Wild-Type. Het1: heterozygous DGCR8 KO line. Scale bar: 200 µm.

4.5. Discussion

To my knowledge, this study presents the first *DGCR8* knock-out out cell model generated in hESCs. The use of the CRISPR/Cas9 genome editing technology allowed the generation of multiple DGCR8 KO cell lines that have undergone homologous recombination in one chromosomal gene copy of the second exon of *DGCR8*. Two lines have been obtained that contains a deletion mutation in the second chromosomal copy of the *DGCR8* gene. However, both mutations were in-frame deletions and thus are not

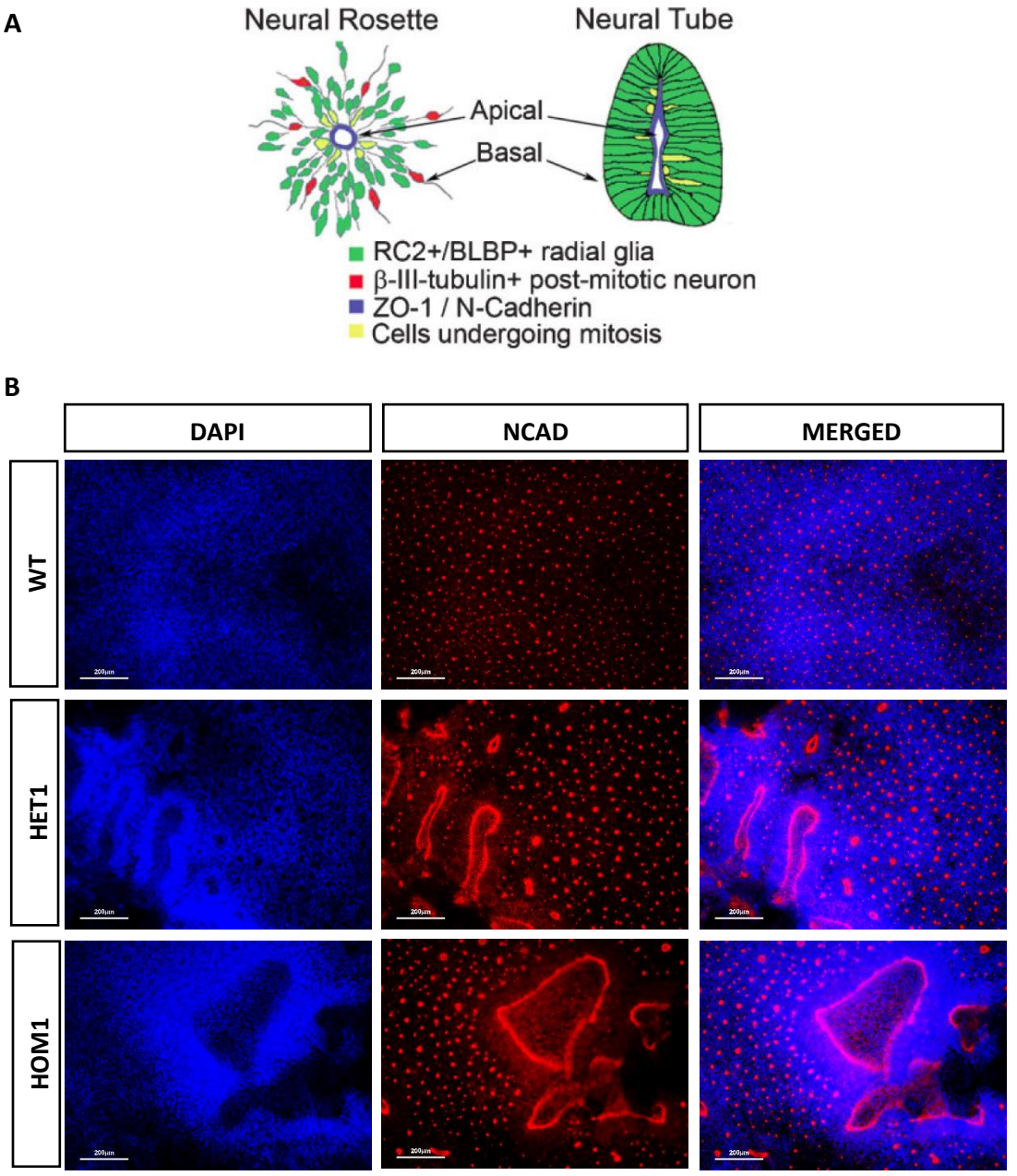


Figure 4-11. Neural rosette formation (A) hESC-derived neural rosettes in vitro bear striking resemblance to the neural tube. Apicobasal polarity is similar between rosettes and the neural tube, with tight junctions (indicated by N-cadherin expression, blue) at the apical surface forming a lumen. Radially arranged progenitors in the rosette, as well as the neural tube, express RC2 and BLBP (green). Figure adapted with permission from John Wiley and Sons: Journal of Cellular Biochemistry, Germain 2010 [500], copyright (2010). (**B**) **Abnormal N-cadherin organisation in *DGCR8* KO lines at the embryonic stem cell stage** (Day 18). DAPI: nuclear counterstains. NCAD: N-cadherin. Scale bar: 200 µm.

expected to cause a termination of translation. No true functional homozygous null mutant has been obtained. Multiple explanations could lead this result. First, it could be due to an insufficient number of clones screened for the mutation, combined with the low rate of effective homology-directed repair obtained. Secondly, *DGCR8* null mutation may

compromise self-renewal and hence reduce the chance of picking and screening such KO lines, for instance by reducing significantly the proliferation rate of the cells as it has been shown in mouse embryonic stem cells [128]. Finally, it is possible that the homozygous knock-out of the *DGCR8* gene is non-viable in human embryonic stem cells. It has however been shown to be viable in mouse embryonic stem cells [128].

The generated KO lines appeared to have an altered DGCR8 protein activity. DGCR8 has the ability to self-regulate its own mRNA levels by binding of the protein to microRNA-like hairpin structure located in the second exon of *DGCR8* mRNA. These structures are then degraded by DROSHA and consequently controlling the protein levels [78,391,414–416]. Here the KO lines have shown to have increased mRNA levels for sequences that are located before the edited locus (Figure 4-6.C, Figure 4-9.A). This difference was significant when targeting the first two exons (before the edited region) in embryonic stem cells for the heterozygous knock-out lines ($p = 0.04$) and the homozygous knock-out line tested ($p = 0.0091$). It was also significant in neural progenitor cells for the homozygous knock-out line ($p = 0.0048$) (Figure 4-8.A, Figure 4-9.C). This result might reflect an altered processing of the stem-loop structure contained in the second exon of DGCR8 mRNA by the microprocessor complex. The increased levels were probably not due to an increased transcription of the DGCR8 mRNA but rather to a reduced degradation of the mRNA due to lower microprocessor proteins levels and/or activity. The Hom1 cells, that contain a 36 bp in-frame deletion in the second allele (and homologous recombinant in first allele), showed a more significant difference (compared to wild-type) when amplifying these mRNA regions situated before the mutated site. This could indicate that the mutation affects the protein function, *i.e.* this second allele might be hypomorphic. However, the difference in mRNA levels when amplifying these regions was not significant when comparing the heterozygous knock-out lines to the homozygous line (data not shown). The presence of this homeostasis loop regulating *DGCR8* mRNA could compensate partially or fully the absence of one chromosomal copy of the DGCR8 gene. For this reason, future studies should determine the protein levels of DGCR8 in these KO lines. However, it is also possible to investigate this indirectly by looking at the levels of other RNAs targeted by DGCR8.

The ability of DGCR8 to bind to other mRNA that contains stem-loop structure -and thus the ability to lead to their degradation through microprocessor complex activity- has also been indirectly tested by looking at the mRNA levels of *NEUROG2*. *NEUROG2*

mRNA contains such structure in its 3' untranslated region and deficits in DROSHA/DGCR8 lead to increased levels of its mRNA [424,425]. Concordantly, the DGCR8 heterozygous and Hom1 KO lines appeared to have increased levels of this mRNA (Figure 4-8.A, Figure 4-9.C). This difference was significant for the heterozygous knock-out lines ($p = 0.00097$) but not for the Hom1 clone ($p = 0.059$). There might have been a non-significant trend in Hom1 clone for an increase of *NEUROD1* mRNA level, another pro-neuronal transcription factor downstream of *NEUROG2* [425]. *PAX6*, a transcription factor interacting with *NEUROG2* [429] and a radial glia marker, did not have altered mRNA levels, and neither did *NESTIN* and *TUJ-1* (neural progenitor cells- and neurons-specific markers). Further analysis of neuronal lineage specific genes by RNA-sequencing could help to further investigate the effect of *DGCR8* knock-out on *NEUROG2* and neural differentiation. Nonetheless, these results suggest a reduced ability of DGCR8 in KO lines to bind and indirectly regulate the mRNA level of *NEUROG2*, containing microRNA-like structures. As NEUROG2 ectopic expression has been shown to induce a rapid neuro-differentiation and neuron maturation [426,428], this could potentially lead to specific cellular phenotypes in *DGCR8* KO lines.

The most studied function of DGCR8 is to bind to primary-microRNAs that are then processed by DROSHA into precursor-microRNA. *DGCR8* KO lines have shown to have increased RNA levels for a small subset of pri-miRNA that have been previously shown to be regulated by the microprocessor complex (pri-miR 16-1 and 2 pri-miRNA which are part of the miR-17~92 gene cluster [415,420]). However, the increase in mRNA levels appeared minimal compared to previously reported results in other cell lines, and the other tested pri-miRNA did not show changes in RNA levels. In particular, pri-miR 185, targeted by DGCR8 and part of the 22q11.2 deletion [78], did not seem altered in the *DGCR8* KO lines. Multiple explanation would be possible, in the hypothesis that the *DGCR8* KO lines are indeed hypomorphic. It could be due to biological variability between the different cell lines tested and between the multiple replicated differentiations used. Secondly, microRNA level changes observed in other cell types might not be reflected in the cell lines used.

A large biological variability has been observed in the replication experiments, as shown by the standard deviations bars in Figure 4-9 and the non-reproducibility of some of the results presented in the preliminary experiments. Nonetheless, the levels of multiple

RNAs have been shown to be significantly altered. These results are indicating that the generated *DGCR8* knock-out lines have an impaired RNA processing ability.

Finally, phenotypic abnormalities have been observed during different runs of neural differentiation of the *DGCR8* KO lines. Notably, large neural-rosette-like structure stained by N-cadherin were present in clones, with a seemingly more robust phenotype in the Hom1 *DGCR8* $^{MUT/-}$ clone that contains a 36 bp deletion on the second allele. Further characterisation of these structures and the cells that compose and surround them is required. In particular, an analysis of earliest time points during differentiation will help to understand if these structures derive from neural rosettes that expand at an increased rate, or that they were the product of fusions of multiple rosettes or other unknown causes.

In summary, multiple *DGCR8* heterozygous and homozygous knock-out cell lines have been generated in human embryonic stem cells. These cells exhibit abnormal molecular and cellular phenotypes. However, future work will be required to further characterise the cellular abnormalities and to elucidate the molecular mechanisms. In particular, mRNA-sequencing will permit to get a better understanding of the transcriptomic changes emerging from the hemizygous loss of *DGCR8*. This will hopefully also give hindsights into the defects in gene regulation occurring in 22q11.2 patients who also suffer from the loss of one copy of the *DGCR8* gene.

5. Transcriptome analysis of DGCR8 knock-out lines

5.1. Introduction

5.1.1. DGCR8 RNA targets and actions

The DGCR8 protein, as a part of the microprocessor complex alongside DROSHA, is an essential component of the canonical microRNA biogenesis pathway (cf. Figure 4-1 Chapter 4 p.82). As microRNAs regulate post-transcriptional expression, DGCR8 is a major player of genetic regulation. However, its involvement extends further due to its ability to bind to other type of RNAs. Kadener *et al.* in 2009 performed a genome-wide identification of microprocessor targets in *D.melanogaster* [430]. It showed in a *Drosha* knock-down model that many RNAs were upregulated similarly to pri-miRNAs. These RNAs were not affected when knocking-down *Dicer*, another protein involved in the canonical microRNA biogenesis pathway. This suggested that the upregulated RNAs were due to *Drosha* knock-down directly and not indirectly through alteration of miRNA levels and their targets. These RNAs have the particularity to contain pri-microRNA-like hairpin secondary structure. They include several mRNAs such as *Dgcr8* itself, and these results have been confirmed in mouse embryonic stem cells [431].

By cross-linking DROSHA or DGCR8 proteins with UV light irradiation or formaldehyde exposure to their RNA partners, multiple studies have been able to sequence and identify these targets [390,432,433]. These studies have shown that the microprocessor complex can bind to coding RNA (mRNA) as well as many subtypes of non-coding RNA, including for instance retrotransposon RNA, long-non-coding RNA (lncRNA), microRNAs (miRNA), small nucleolar RNA (snoRNA), ribosomal RNA (rRNA) and mitochondrial RNA (mtRNA). snoRNA are a class of non-coding RNA involved in guiding chemical modification of other RNA molecules (such as rRNA) [434], and are regulated by DGCR8 independently of DROSHA [390]. It has been suggested that the DGCR8 protein can form a complex with the nuclear exosome and recruit it to induce the degradation of double stranded structured DNA such as these snoRNAs or the telomerase RNA component, essential for telomere replication [435].

There is evidence that DROSHA and DGCR8 can affect gene expression in multiple ways. The first one is the most studied, through the regulation of microRNAs and thus mRNA targets levels. Secondly, it can directly alter the level of mRNAs that contain

exonic microRNAs by cleaving the secondary structure. This mechanism has been demonstrated for *Dgcr8* mRNA [390,391,414,415,431], *Neurog2* mRNA [424,425] , *Aurkb* [432] mRNA as well as *Hoxa7*, *Dlg5* and *Snx12* mRNAs [390]. The microprocessor complex can also regulate gene expression independently of its RNA-cleavage function. It has been shown by chromatin immunoprecipitation analysis that both DROSHA and DGCR8 protein can bind to the promoter-proximal regions of large number of genes [436]. DROSHA interacts with the RNA polymerase II protein and this activity correlates with gene expression levels [436].

The two main proteins of the microprocessor complex (DROSHA/DGCR8) also have a role in splicing events. DGCR8 regulates the relative levels of alternatively spliced transcripts by binding to specific exons that can in turn (but not necessarily) be cleaved [390]. It has been for instance shown that the exon 5 of the *eIF4H* gene, which contains a predicted hairpin structure, is cleaved by DROSHA [437]. Microprocessor-dependent alternative splicing can also occur due to the presence of hairpin structures located across exon-intron junctions [438]. DGCR8 and DROSHA are found within the supraspliceosome complex, and it has been shown that the activity of either one of the microprocessor complex or spliceosome is negatively correlated with the activity (and protein levels) of the other [439].

5.1.2. 22q11.2 deletion syndrome transcriptome and relation to DGCR8

5.1.2.1. Mouse models

The different processing activity of the microprocessor on a wide range of RNA subtypes suggests that the hemizygous deletion of *Dgcr8* in knock-out models or in the context of 22q11.2 deletion syndrome is likely to lead to transcriptome changes. Stark *et al.* generated in 2008 the *Df(16)A*$^{+/-}$ mouse model that carry a hemizygous chromosomal deficiency spanning a region syntenic to the 1.5 mb A-B deletion at chromosome 22q11.2 in humans [78]. Analysis of the *Df(16)A*$^{+/-}$ transcriptome in the prefrontal cortex (PFC) and the hippocampus revealed changes at both mRNA and microRNA levels. Only a subset of the primary-microRNAs tested in the study were shown to have increased levels in the mutant compared to the wild-type mice and only some of their mature microRNAs were downregulated. Affected pri-miRNA and miRNAs were found to have similarly altered levels in *Dgcr8*-deficient mice, suggesting that this protein was the cause of these differences. In the *Df(16)A*$^{+/-}$ mice, more gene transcripts were significantly differentially

expressed in the PFC than the hippocampus, but a small subset of genes were dysregulated in both tissue type, including genes outside the 22q11.2 deleted region. Gene ontology analysis of dysregulated genes suggested that those affected in the PFC are related to energy metabolism while those affected in the hippocampus were related to synaptic functions. Interestingly, looking for miRNA target sites (seed sites) within the 3'untranslated region (UTR) of the dysregulated genes identified an enrichment of seed sequences in upregulated genes. This suggests that part of the alteration of the transcriptome in the 22q11.2DS mice model is indeed due to decreased levels of DGCR8-dependent microRNAs. Further analysis of this model showed that one microRNA in particular, miR-185, represses *Mirta22*. This is particularly relevant because miR-185 is a microRNA whose sequence is located within the 22q11.2 deletion, and is also processed by the microprocessor complex, thus resulting an important reduction in 22q11.2 mice [176]. This reduction of miR-185 leads to greater *Mirta22* levels which in turn affects neuronal dendritic and spine growth in this 22q11.2DS mouse model [176]. Interestingly, it has been shown that some of the genes with lower hippocampal mRNA levels in *Dgcr8*$^{+/-}$ mice (compared to wild-type) have also been shown previously to be downregulated in schizophrenia patients [398]. This further suggests the importance of studying DGCR8-related gene regulation and its impact on the transcriptome to understand schizophrenia aetiology in the 22q11.2 deletion syndrome.

5.1.2.2. 22q11.2 DS patients

Changes in microRNA levels that have been observed in mouse embryonic stem cell *Dgcr8* knock out, mouse *Dgcr8* knock-out or 22q11.2 mouse models have been similarly reported in peripheral leucocytes from 22q11.2 deletion syndrome patients [129]. This includes miR-185, which was the most significantly downregulated mature miRNA reported in the study. These patients were also found to have significantly decreased levels of *DGCR8* mRNA and protein. Microarray analysis of gene expression of peripheral blood mononuclear cells (PBMCs) of 22q11.2 patients compared to individuals without this CNV showed a dysregulation of genes implicated in functional networks that can be linked to the disorder, such as the nervous system development or cardiovascular system development [440]. A potential bias of both studies is that 22q11.2 deletion syndrome patients have been reported to have lower T-cell numbers [43,44]. As a consequence, it might result in apparent changes in gene expression caused by different

cell-type compositions between individuals with or without the CNV (cf. Chapter 2 (p.31) for further discussion on this subject in the context of DNA methylation).

Jalbrzikowski *et al.* in 2015 have shown that 65% of the differentially expressed genes in peripheral blood from 22q11.2 deletion syndrome patients are expressed in the brain [441]. GO analysis showed that the significant genes were enriched in neuron-related ontology terms, while gene network IPA analysis showed an association with cellular development, growth and proliferation as well as axon guidance signalling. One great strength of the study was that they performed a comparison of the transcriptomes of 22q11.2 deletion syndrome patients with or without neuropsychiatric disorders. First, this allowed an analysis of the gene expression differences between patients that develop different phenotypes. Moreover, while the issue was not addressed in the study, it should reduce the cell-type composition bias. In 22q11.2 deletion syndrome patients, psychosis status was associated with differential expression of genes related to mitochondrial organisation and transport (GO terms) as well as embryonic development, cellular assembly and cellular organisation (canonical pathways). Differentially expressed genes depending on autism spectrum disorder status in 22q11.2 deletion syndrome were associated with cell morphogenesis in differentiation (GO term) and immune response (canonical pathway).

5.1.2.3. Stem cell models

The development of the Induced Pluripotent Stem Cell (iPSC) technology has been a major step forward in the investigation of neurological and neuropsychiatric disorders. This technique allows the reprogramming of differentiated cells such as skin fibroblast into pluripotent cells that can be differentiated into many other cell types (as reviewed in Shi 2017 [442]). One first advantage of this technique is to be able to study the impact of known human mutations in cell types and tissues that might be difficult to access, such as the brain. Stem cell differentiation protocols have been developed to obtain different types of neurons, for instance cortical neurons, interneurons or dopaminergic neurons (as reviewed by Mertens *et al.* 2016 [186]). This can facilitate the investigation of disease mechanisms in relevant cell types. The second main advantage of the hiPSC technology is to study genetic environments that are too complex to be generated with the current existing genome editing technologies. Common variants that leads to an increased risk in schizophrenia in the general population or in 22q11.2 deletion syndrome patients (cf.

Chapter 2 (p.31)) are numerous, and recreating particular combinations of alleles in stem cell lines would be extremely strenuous. Despite recent advances in the CRISPR/Cas9 genome editing system to generate large CNVs [443], the size of the 22q11.2 deletion makes it difficult to re-create in human embryonic stem cells. No 22q11.2 deletion has been generated yet in embryonic stem cells to my knowledge.

The first study of gene expression in neurons differentiated from hiPSCs derived from 22q11.2 deletion syndrome patients with schizophrenia was performed in 2011 by Pedrosa *et al.*[444]. It revealed that the expression of pluripotency markers *Oct4* and *Nanog* are retained for an increased period of time after induction of differentiation; similar results have been demonstrated in *Dgcr8* knock-out mouse embryonic stem cells [128]. Microarray transcriptome analysis in hiPSC-derived neurons carrying a 22q11.2 deletion revealed that at both neural progenitor stage (day 10) and more mature neurons (day 32) there was an enrichment of genes involved in Gene Ontology (GO) terms related to neuron morphogenesis as well as cell cycle control and proliferation. There were also strong enrichments in biological pathways related to neurological diseases, psychological disorders and the cell cycle. A more recent analysis by the same group used RNA-sequencing and an increased number of hiPSCs-derived cell lines from 22q11.2DS patients with psychotic disorders and controls to further investigate gene expression changes related to the deletion [445]. The top GO terms that were enriched for up-regulated genes in 22q11.2DS neurons were related to apoptosis and the immune response, while for down-regulated genes it was related to cell cycle as well as the glutamate metabolic process, neurotransmission and cytoskeletal organisation terms. One of the canonical pathways reported as enriched was ERK/MAPK, which is consistent with what has been shown in Van Beveren's study in PBMCs from 22q11.2 DS patients [440].

Another independent study of gene expression (microarrays) of neurons derived from 22q11.2DS hiPSCs showed relatively similar results, with an enrichment of downregulated genes for cell-cycle-related pathways and an enrichment of up-regulated genes for MAPK-signaling and neurotransmission-related terms [446]. MicroRNA-sequencing of neurons (mainly glutamatergic and GABAergic) derived from 22q11.2DS hiPSCs with psychotic disorders revealed 45 differentially expressed microRNAs compared to non-22q11.2DS non-schizophrenic controls [419]. It is possible that this could be due to the significant decrease of *DGCR8* expression in these cells due to the haploinsufficiency of the gene. However, only 6 of them reached genome-wide

significance, of which 4 microRNAs are mapped within the 22q11.2 deletion (including miR-185). Interestingly, despite not reaching genome-wide significance, two thirds of the differentially expressed microRNAs were upregulated, which is unlikely to be due to a direct effect of the *DGCR8* deletion. The top disease enrichment (DAVID functional annotation) for predicted targets of these upregulated microRNAs was schizophrenia. Bundo *et al.* have also reported that neurons derived from 22q11.2DS hiPSCs have an increased retrotransposition of long interspersed nucleotide element-1 (L1) [447]. They also revealed an increase of L1 content in brain tissue from schizophrenia patients and that insertion of these elements was preferential into genes related to this neuropsychiatric disorder and to synaptic function. This is of particular interest in the context of 22q11.2 as the microprocessor complex has been shown to bind to mammalian retrotransposons mRNAs such as L1 elements mRNAs (interaction with DGCR8) and control their levels via DROSHA-mediated cleavage [448].

To date there is no transcriptomic data from *Dgcr8* knock-out human stem-cell derived neuronal progenitor cells (or other neural lineage cells such as neurons). MicroRNA-sequencing has however been performed in neurons from mouse embryos (E13.5) that have a brain-specific conditional knock-out of *Dgcr8* [449]. This study showed the expected decrease of microRNA reads, with a stronger effect in the *Dgcr8* homozygous knock-out mice than in *Dgcr8* heterozygous knock-out mice. *Dgcr8* KO mouse embryonic stem cell also shows a smaller depletion of microRNAs in heterozygous knock-outs than in null mutants [450]. Similarly, for mRNA expression, only a small number of differentially expressed genes were observed in heterozygous cell lines while between around 2000 to 3000 mRNA levels were significantly altered in homozygous lines. Studies of *Dgcr8* null mutants mESCs have shown that the resulting miRNA deficiency alters expression of genes involved in pluripotency, promoting toward a pluripotent stem cell ground state [298] that favours stem cell self-renewal [451]. In particular, DGCR8-dependent microRNAs from the miR-290/302 family have been shown to control this state by repressing genes from the AKT pathway and upregulating genes from the MEK/ERK pathway [452].

5.2. Aims of the chapter

The work described in this chapter set out to conduct a novel investigation of the effect of *DGCR8* depletion on gene expression during neural differentiation from human embryonic stem cells. The goals of the project are the following:

- Conduct whole transcriptome gene expression analysis by RNA-sequencing and determine which transcripts are differentially expressed in DGCR8 depleted lines.

- Determine which biological pathways are enriched for differentially expressed transcripts.

- Investigate the presence of transcripts that are similarly differentially expressed in *DGCR8*$^{+/-}$ hESCs- and 22q11.2 IPSCs-derived neurons.

5.3. Methods

5.3.1. Cell lines

6 human embryonic stem cell (hESCs) lines were used in this work: WT, WTcrispr1, WTcrispr2, Het1, Het2, Het3. They are further described in Chapter 4 Sections 4.4.1 (p.91) and 4.4.3 (p.98). Briefly, the WT line is a human embryonic stem cell line (H7 line) from which all the other hESCs lines were derived. Het1, Het2, Het3 are 3 distinct DGCR8 $^{+/-}$ heterozygous knock-out lines generated with the CRISPR/Cas9 technology. WTcrispr1 and WTcrispr2 are two wildtype control lines that went through the same CRISPR/Cas9 protocol but for which the genome editing was not successful at the targeted locus. WTcrispr1 results from the same experiment than the Het1/2/3 lines (targeting of *DGCR8*). WTcrispr2 has been provided by Dr. Lucia Cardo (Cardiff University) and results from an independent CRISPR/Cas9 targeting study.

2 human induced pluripotent stem cell (hiPSC) lines were used in this study: 22q11.2DS hiPSC line and WT hiPSC line. The biopsies used to generate the hIPSC lines were provided by the National Centre for Mental Health (NCMH). The 22q11.2DS hiPSC was derived from a 27-year old female affected by neuropsychiatric disorders (Schizophrenia, Autism spectrum disorders, major depressive disorder (recurrent, sever), learning difficulties, panic disorder and agoraphobia). 22q11.2 deletion was maternally inherited. The patient was taking antipsychotic medication and the presence of the deletion in the 22q11.2DS hiPSC was confirmed by SNP array (mapped to Chr22:18892575-21452237 region (LCR A to LCR D), build 37/hg19). The control line WT hiPSC was derived from a 27-year-old male individual with no known history of psychotic disorder. The presence of the 22q11.2 deletion was not assessed in this individual. Consent forms (DEFINE/ECHO consent form, NCMH cell biology consent form) were signed by both subjects. Both hiPSC lines were generated by Dr. Craig Joyce

(Cardiff University). Dermal fibroblasts derived from skin punch biopsies were reprogrammed into hiPSCs using the CytoTune-iPS 2.0 Sendai Reprogramming Kit (ThermoFisher Scientific). Cell pluripotency of hiPSCs was confirmed by immunocytochemistry using antibodies against pluripotency markers (Nanog, Oct3/4 and Tra1-81).

5.3.2. Neural differentiation and RNA extraction

Samples used for RNA-sequencing were from neural progenitor cells extracted at day 18 of a dual-SMAD signalling inhibition cortical neuron differentiation protocol as described in Chapter 4 Section 4.3.2 (p.87). Total RNA was extracted with mirVana miRNA Isolation Kit (ThermoFisher Scientific) according to manufacturer's protocol. For each sample, cells from two wells of a twelve-well plate were lysed and pooled together to reduce intra cell-line sample variability. RNA integrity was assessed with Bioanalyzer RNA Nano Kits (Agilent) on an Agilent 2100 Bioanalyzer instrument by Dr. Amanda Redfern (Central Biotechnology Services, Cardiff University). All samples had a RNA Integrity Number RIN > 8.6 (mean = 9.46, s.d. = 0.28).

Two independent experiments have been performed. For each experiment, all cell lines were differentiated simultaneously.

- Experiment 1: 8 cell lines: WT, WTcrispr1, WTcrispr2, Het1, Het2, Het3. All samples were extracted at day 18. 3 biological replicates have been extracted for each cell line.

- Experiment 2: 4 cell lines: WT, Het1, 22q11.2DS hiPSC, WT hiPSC. This experiment was performed in collaboration with Matthieu Trigano (Cardiff University) (differentiation and RNA extraction only). 2 other cell lines were used (a hESC *ZDHHC8*$^{+/-}$ knock-out line and its isogenic control). However, results from these two cell lines were analysed by Matthieu Trigano and are not discussed in this thesis. WT and Het1 cell lines were extracted at day 18, other cell lines were extracted at day 15 to account for the faster differentiation rate of these cell lines that have a different background than WT/Het1 lines. 6 biological replicates have been extracted for each cell line.

5.3.3. RNA-seq data acquisition

Sample preparation and sequencing was performed by Joanne Morgan (NGS Co-ordinator, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University). RNA was quantified with QubitRNA High Sensitivity kits (ThermoFisher). Sequencing libraries for mRNA transcriptome analysis were prepared using KAPA mRNA HyperPrep kits (Kapa Biosystem) from 1 µg of total RNA per sample. The standard protocol was followed, with the following options/modifications: fragmentation was performed for 6 min at 94°C; 8 cycles of PCR were used for the library amplification step; after amplification, samples were cleaned for two rounds using Agencourt AMPure XP (Beckman Coulter). Libraries were quantified using QubitRNA High Sensitivity kits (ThermoFisher) and sized using High Sensitivity DNA kits (Agilent). Libraries were pooled in equimolar amounts (10 nM). The experiment 1 was pooled as a 24-plex; the experiment 2 was constituted of two 18-plex pools (with equal distribution of samples across the two pools). Clustering and sequencing was carried out as per standard Illumina protocols. Library pools were diluted to 3 nM for clustering procedure and spiked with 1% PhiX control genome. The prepared libraries were sequenced using Illumina's Hiseq 4000 with four each lane for each experiment which generated 2 x 75 bp paired-end reads (4 lanes for each sample for Experiment 1, 24 samples per lane; 2 lanes for each sample for experiment B, 18 samples per lane). This resulted in an average of 57.3 million read pairs generated per samples for Experiment 1 (s.d. = 6.5 million reads) and an average of 38.1 million sequenced read pairs per samples for Experiment 2 (s.d. = 5.2 million reads).

Alignment of RNA-seq reads was performed by Daniel Cabezas De La Fuente (Cardiff University) using a script developed and tested by Daniel Cabezas De La Fuente and Dr. Robert Andrews (Cardiff University). Briefly, data was pre-processed with the Trimmotic software to trim adaptor sequences and remove poor-quality bases [453]. Bases with a Phred quality score Q < 30 were eliminated. Data quality control was performed with FastQC before and after trimming [454]. Data was aligned to the hg38 reference genome using the STAR aligner software version [455].

5.3.4. Data analysis

Data analysis was performed in R (Microsoft R Open 3.4.0 (Microsoft), based on R-3.4.0 (R Statistics)) using a modified version of a script developed by Daniel Cabezas De

La Fuente and Dr. Robert Andrews (Cardiff University). Genes with less than 10 counts over all samples were removed from the analysis. Differential gene expression analysis was performed with the *DESeq2* R package [456,457]. Principal component analyses were conducted after performing a regularized log transformation (function *rlog*, *DESeq2*). Functional enrichment analyses (Gene Ontology (GO) and KEGG pathways) were performed with the *clusterProfiler* R package [458] (*enrichGO* and *enrichKEGG* functions). Samples from the same experiments (experiment 1 or experiment 2) were all normalised together and then either analysed together or divided in subsets for differential expression analysis. Due to the gender difference between the two hiPSCs lines used, only the significant differentially expressed genes obtained when comparing 22q11.2 DS hiPSCs vs WT hESCs (both female cell lines) were retained in the list of genes obtained when comparing 22q11.2 DS hiPSC (female) versus control hiPSCs (male).

5.4. Results

5.4.1. Experiment 1: differentially expressed genes in *DGCR8*$^{+/-}$ neural progenitor cells.

RNA-sequencing was performed with DGCR8 heterozygous knock-out human embryonic stem cells and control embryonic stem cells that were differentiated in parallel into neural progenitor cells (NPCs). 3 *DGCR8*$^{+/-}$ KO lines were used, as well as three wild-type (WT) control lines (including two lines that were processed in the same way as the CRISPR/Cas9 edited cell lines but for which no mutation was been detected at the targeted locus). Comparing the three KO lines to the three WT lines identified 1744 transcripts that were differentially expressed with a FDR-corrected p-value $p < 0.01$. This included 979 transcripts that were upregulated and 765 transcripts that were downregulated.

Principal component analysis (PCA) of the RNA-sequencing data revealed a high heterogeneity in the transcriptome of the different cell lines tested, in particular between the wild-type control lines. Figure 5-1.A shows the PCA results for the top 500 most variable genes between samples (variance of normalised read counts between sample per gene). The three KO lines cluster together relatively well, however the three WT cell lines appear to cluster independently. One Wild-Type line, WTcrispr1, appears to cluster with the KO lines. This line is a wild-type sister clone of the KO lines and has been derived from the same CRISPR/Cas9 editing process. This could therefore be due to off-target

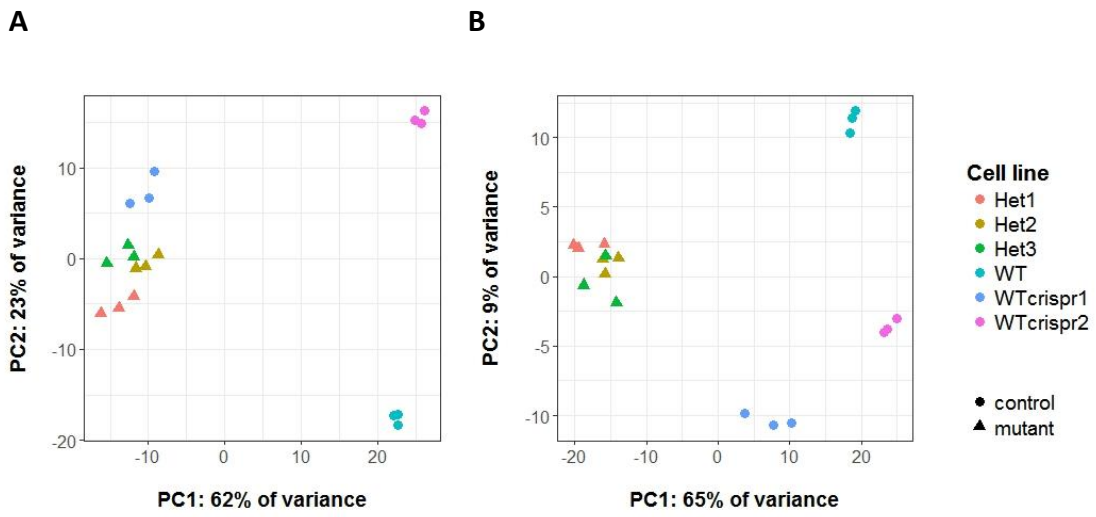


Figure 5-1. Principal component analysis (PCA) of RNA-sequencing results and variance explained by the first two principal components. (**A**) PCA based on the top 500 most variables genes. (**B**) PCA based on the top 500 most significant genes (Het1/2/3 lines versus WT lines). PCAs are based on regularized log count data. Het1/2/3: *DGCR8* $^{+/-}$ heterozygous knock out lines; WT/WTcrispr1/WTcrispr2: wild-type control lines.

effects of the CRISPR/Cas9 process, or it could reflect the fact that they are derived from the same parental cells (WT line is the same parental line but from a different aliquot of cells.). Figure 5-1.B shows the PCA results for the top 500 most significant genes when comparing the KO lines to the WT lines. It shows that despite focusing on the genes that differentiate the most between KO and WT neural progenitor cells, there still remains variability in expression between the wild-type lines.

5.4.1.1. Differential expression analysis

The heterogeneity in the transcriptome of the different control lines could induce bias in the analysis and lead to false gene discovery. In order to determine a list of high-confidence genes that were differentially expressed, the experiment was divided into 3 sub-analyses. Each analysis compared independently one of the KO line versus one of the WT control line (cf. Figure 5-2.A/B). While this reduces the power of each analysis due to the small number of samples used, it was based on the assumption that any genes that

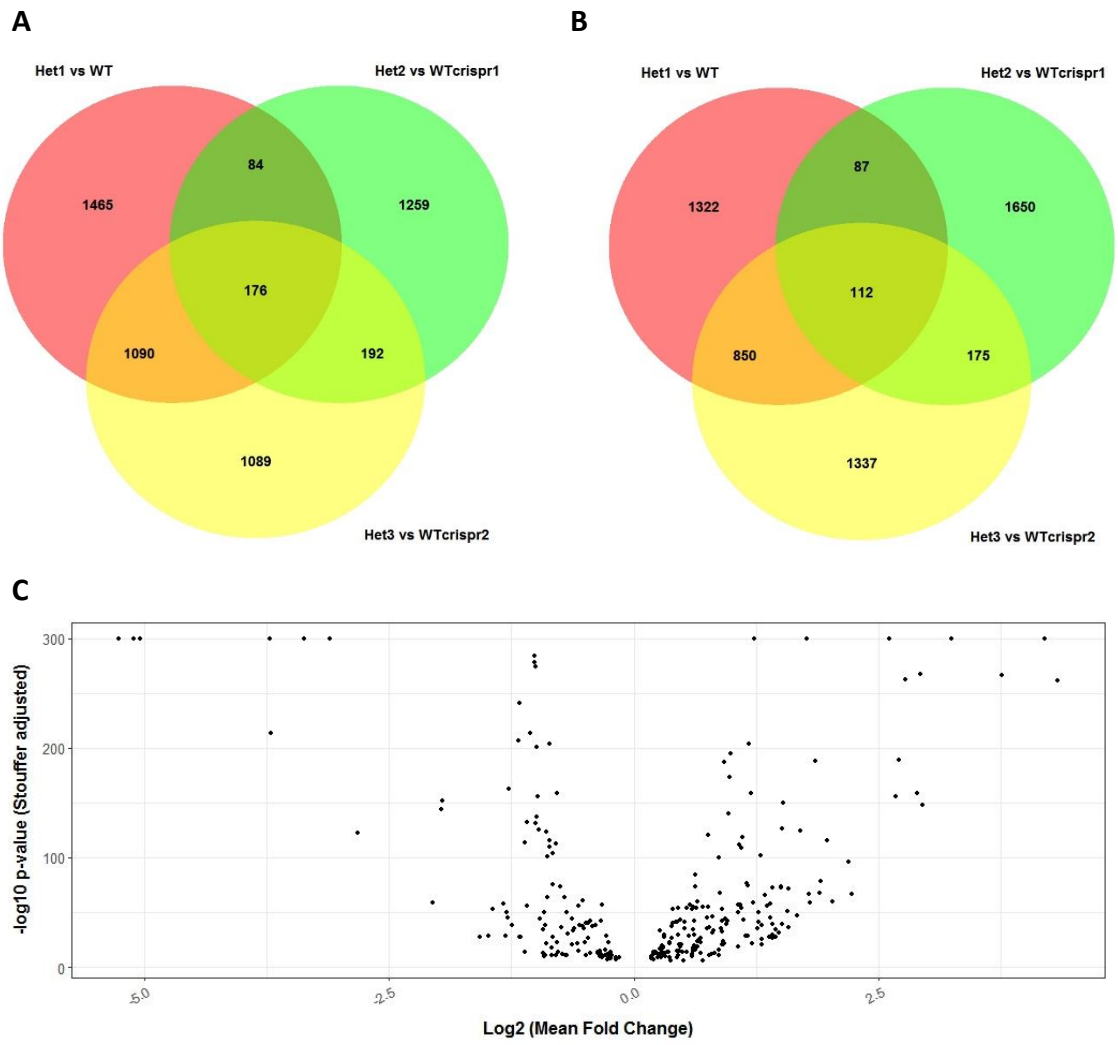


Figure 5-2. Analysis of high-confidence genes differentially expressed in *DGCR8*$^{+/-}$ neural progenitor cells. 3 independent differential expression were performed with different *DGCR8*$^{+/-}$ lines and Wild-Type lines (**A**) Upregulated genes in *DGCR8*$^{+/-}$ lines. (**B**) Downregulated genes in DGCR8$^{+/-}$ lines. (**C**) Volcano plots of genes significantly differentially expressed in all three independent comparisons. Fold change: mean fold change over the three experiments; p-values: combined p-values for the three independent experiments (Stouffer's method). P-value < 1e^{-300} are set to p = 1e^{-300}. Het1/2/3: *DGCR8*$^{+/-}$ heterozygous knock out lines; WT/WTcrispr1/WTcrispr2: wild-type control lines.

are differentially expressed in all 3 separate analyses are more likely to represent a true signal and be potentially biologically relevant. This analysis resulted in 288 genes that were differentially expressed in all 3 independent analyses. 176 of these genes were upregulated in *DGCR8* KO NPCs (Figure 5-2.A) and 112 were downregulated (Figure 5-2.B). Their fold-change values were averaged, and their p-values were combined with the Stouffer method [318]. Figure 5-2.C represent these value for each of genes.

A

GO term	GO type	Adjusted p-value
synapse organization	BP	1.4E-02
C21-steroid hormone biosynthetic process	BP	1.4E-02
muscle tissue development	BP	2.9E-02
regulation of binding	BP	2.9E-02
synapse assembly	BP	2.9E-02
membrane biogenesis	BP	4.2E-02
proteinaceous extracellular matrix	CC	4.3E-02

B

GO term	GO type	Adjusted p-value
proteinaceous extracellular matrix	CC	8.3E-05
extracellular matrix	CC	1.1E-03
heparin binding	MF	1.1E-03
sulfur compound binding	MF	3.4E-03
glycosaminoglycan binding	MF	3.6E-03
synapse organization	BP	6.7E-03
C21-steroid hormone biosynthetic process	BP	1.4E-02
connective tissue development	BP	1.4E-02
regulation of striated muscle tissue development	BP	1.4E-02
regulation of muscle organ development	BP	1.4E-02
negative regulation of steroid biosynthetic process	BP	1.4E-02
negative regulation of steroid metabolic process	BP	1.4E-02
skeletal system development	BP	1.4E-02
regulation of muscle tissue development	BP	1.4E-02
transmembrane receptor protein serine/threonine kinase signaling pathway	BP	2.2E-02
glucocorticoid biosynthetic process	BP	2.2E-02
regulation of ketone biosynthetic process	BP	2.2E-02
aldehyde biosynthetic process	BP	2.2E-02
axon guidance	BP	2.9E-02
neuron projection guidance	BP	2.9E-02
synapse assembly	BP	2.9E-02
endocardial cushion development	BP	2.9E-02
collagen metabolic process	BP	2.9E-02
ossification	BP	2.9E-02
multicellular organism catabolic process	BP	2.9E-02
cell morphogenesis involved in neuron differentiation	BP	2.9E-02
hindbrain development	BP	3.0E-02
regulation of hormone biosynthetic process	BP	3.6E-02
osteoblast differentiation	BP	3.6E-02
collagen trimer	CC	3.9E-02
glucocorticoid metabolic process	BP	4.0E-02
negative regulation of DNA binding	BP	4.7E-02
transforming growth factor beta receptor signaling pathway	BP	4.7E-02
developmental growth involved in morphogenesis	BP	4.7E-02
cellular hormone metabolic process	BP	5.0E-02

Table 5-1. Gene Ontology gene-set enrichment analysis of high-confidence genes differentially expressed in *DGCR8*$^{+/-}$ neural progenitor cells. (**A**) All significant genes. (**B**) Upregulated genes only. Note: No significant enrichment for downregulated genes only. Differentially expressed genes are defined as genes with an FDR corrected p-value < 0.01. Significantly enriched terms have an adjusted p-value < 0.05 (Benjamini-Hochberg procedure). BP: biological process; CC: cellular component; MF: molecular function.

5.4.1.2. Biological pathway enrichment analysis

Functional enrichment analysis was performed on this list of 288 significantly differentially expressed genes to determine if they were enriched for specific Gene

KEGG Pathway	Adjusted p-value
Hippo signaling pathway	1.6E-02
TGF-beta signaling pathway	1.6E-02
Axon guidance	2.9E-02
Amoebiasis	4.2E-02
ECM-receptor interaction	4.3E-02

Table 5-2. KEGG pathway gene-set enrichment analysis of high-confidence genes upregulated in *DGCR8*$^{+/-}$ neural progenitor cells. Only upregulated genes are used in the enrichment analysis. Note: No significant enrichment for downregulated genes or all genes together. Differentially expressed genes are defined as genes with an FDR corrected p-value < 0.01. Significantly enriched pathways have an adjusted p-value < 0.05 (Benjamini-Hochberg procedure).

Ontology (GO) terms or KEGG pathways. This analysis identified multiple biological process GO terms that were significantly enriched ($p < 0.05$) in genes that were differentially expressed with p-value < 0.01 (Table 5-1.A). In particular, the most significant biological process GO terms was Synapse organisation ($p = 0.014$). Restricting to genes that were upregulated leads to the same most significant Biological Process term (Table 5-1.B). No GO terms were significant when the test was restricted to down-regulated genes. This suggests that the detected enrichment was mainly being driven by genes that have greater expression in *DGCR8*$^{+/-}$ KO lines than in the WT lines. No KEGG pathway reached significance when testing for enrichment of all genes or downregulated genes. Restricting to upregulated genes shows an enrichment in the Hippo and TGF-beta signalling pathways ($p= 0.016$ and $p=0.016$ respectively) as well as Axon guidance ($p = 0.029$) (Table 5-2).

5.4.2. Experiment 2: DGCR8$^{+/-}$ and 22q11.2 deletion shared transcriptome changes

A second RNA-sequencing experiment was performed to investigate possible similarities between changes in expression due to *DGCR8* heterozygous deletions and the ~3 Mb deletion at 22q11.2 (which also spans DGCR8). For this, 4 cell lines were differentiated in parallel into neural progenitor cells: one *DGCR8*$^{+/-}$ KO human embryonic stem cell line (hESCs), one Wild-Type control hESCs, one induced pluripotent stem cell (hiPSC) line derived from a patient with the 22q11.2 deletion syndrome and one control hiPSC. NPCs differentiated from hiPSCs or hESCs were analysed separately to find differentially expressed genes due to *DGCR8* KO or 22q11.2 deletion independently, and the results were compared. 457 genes were significantly

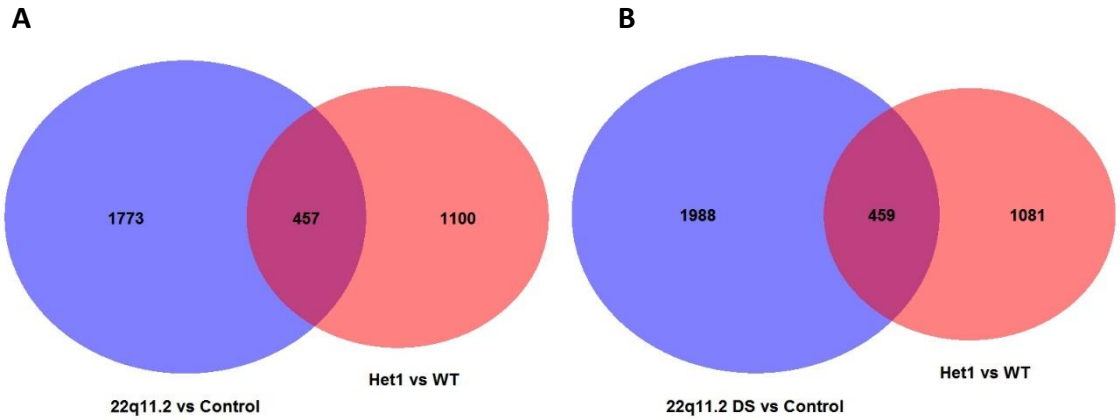


Figure 5-3. Overlap between genes differentially expressed in *DGCR8*$^{+/-}$ KO neural progenitor cells and 22q11.2 DS neural progenitor cells (NPCs). (**A**) Upregulated genes (**B**) Downregulated genes. 22q11.2: cell line with the 22q11.2 deletion; Control: control line without the deletion; Het1: *DGCR8*$^{+/-}$ heterozygous knock out lines; WT: wild-type control lines.

differentially upregulated in both *DGCR8* KO NPCs and 22q11.2DS NPS (FDR-corrected p-value < 0.01) while 459 genes were significantly differentially downregulated (Figure 5-3.A and B respectively). Gene-sets enrichment analysis of genes that were upregulated or downregulated in both 22q11.2 DS NPCs or DGCR8$^{+/-}$ NPCs are detailed in Supplementary table 5-1.

5.4.3. Investigating high-confidence genes affected by both the *DGCR8* deletion in *DGCR8* KO and 22q11.2DS

Due to the variability of differentially expressed genes due to *DGCR8* heterozygous KO depending on which control line is used (Experiment 1, Figure 5-2.A/B), I decided to investigate which of the high-confidence genes determined in the first experiment were replicated in the second experiment (note: the two lines Het1 and WT used in the second experiment were also used in the first experiment, but the results were from different experiments of neural differentiation and RNA-sequencing). Of the 288 high-confidence genes determined the Experiment 1, 99 were replicated in Experiment 2 (Figure 5-4.A/B). Of these, 39 were upregulated and 60 were down-regulated. 63 genes were also significantly differentially expressed in NPCs differentiated from 22q11.2 DS hiPSCs; 15 were upregulated, 48 were downregulated.

Mapping these genes to their respective chromosomal locations revealed that around 40% (25 genes) were mapped to the same chromosomal location on chromosome 17

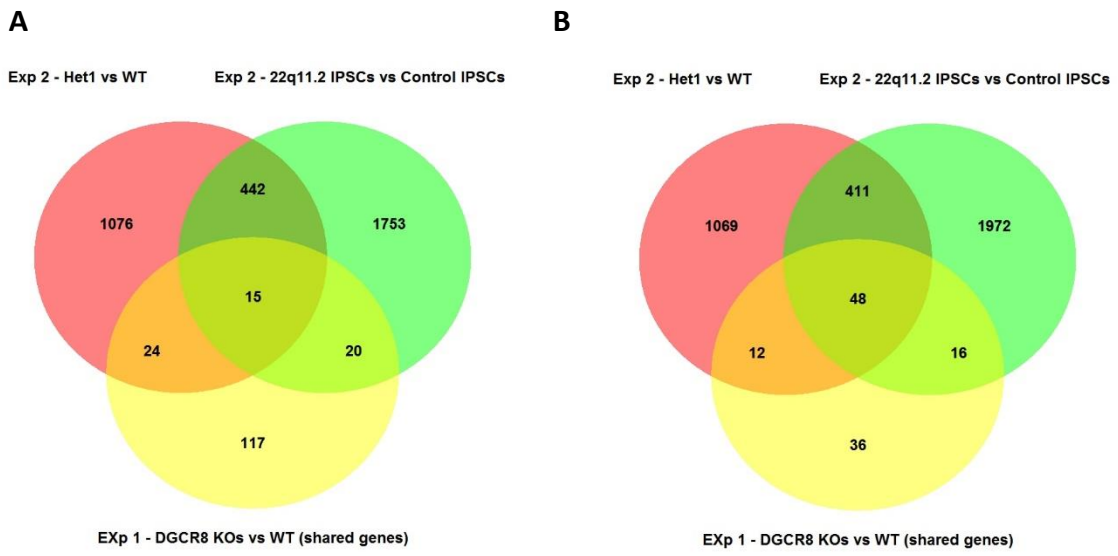


Figure 5-4. High-confidence genes differentially expressed in *DGCR8*$^{+/-}$ and 22q11 DS neural progenitor cells. (**A**) Upregulated genes (**B**) Downregulated genes (**C**) (Next page) Chromosomal locations of genes that are differentially expressed in *DGCR8*$^{+/-}$ KO NPCs (experiment 1 and 2) as well as 22q11.2DS NPCs. Gene ID: ENSEMBL gene ID. 4 Gene ID were not recognised by the R package used and are not represented.

(Figure 5-4.C). It encompasses around 0.4 Mb and include genes from ZBTB4 to AC104581.5. This genomic region corresponds to the cytoband 17p13.1. The average fold-change for this set of significant genes located within this region is FC = -1.92 (s.d. = 0.25) when comparing DGCR8$^{+/-}$ lines to controls in the Experiment 1, FC = -2.09 (s.d. = 0.35) when comparing the DGCR8$^{+/-}$ line to its control in the Experiment 2 and FC = -2.1 (s.d. = 0.44) when comparing the 22q11.2 DS line to its control in the Experiment 2. The normalised read counts for all genes located at the region 17p13.1 (and included between ZBTB4 and AC104581.5) for each of these comparisons and their significance levels are represented in Supplementary figure 5-1.

The pairwise comparisons overlaps between the three groups shown in Figure 5-4 are statistically significant for both downregulated and upregulated genes, as presented Figure 5-5. These overlap remains significant after removing the genes located within the 17p31.1 region (for the intersect of the three groups, $p = 1.17 * 10^{-13}$ for upregulated genes and $p = 7.71 * 10^{-35}$ for downregulated genes).

C

Gene name	Chromosomal location	Cytoband	Direction
FPGT	Chr1:74198212-74234086	1p31.1	Downregulated
LINC01139	Chr1:238480384-238486023	1q43	Upregulated
PLCL1	Chr2:197804702-198572581	2q33.1	Downregulated
PTH1R	Chr3:46877746-46903799	3p21.31	Upregulated
OCIAD2	Chr4:48885019-48906937	4p11	Downregulated
NAP1L5	Chr4:88695915-88698235	4q22.1	Downregulated
TICAM2	Chr5:115578650-115602479	5q22.3	Upregulated
PCDHB10	Chr5:141192353-141195642	5q31.3	Upregulated
TFAP2A	Chr6:10393186-10419659	6p24.3	Downregulated
ZNF311	Chr6:28994785-29005316	6p22.1	Upregulated
PTCHD4	Chr6:47878028-48068689	6p12.3	Downregulated
ZNF483	Chr9:111525159-111577844	9q31.3	Downregulated
ECHDC3	Chr10:11742366-11764070	10p14	Upregulated
SPAG6	Chr10:22345445-22454224	10p12.2	Upregulated
ZNF22	Chr10:45000475-45005326	10q11.21	Downregulated
CEP164P1	Chr10:45002222-45076066	10q11.21	Downregulated
LINC01515	Chr10:65570338-65768835	10q21.3	Downregulated
TRIM68	Chr11:4598672-4608259	11p15.4	Downregulated
TRIM6	Chr11:5596109-5612958	11p15.4	Upregulated
ALKBH3	Chr11:43880811-43920266	11p11.2	Downregulated
AP000769.1	Chr11:65455258-65466720	11q13.1	Upregulated
AC092490.1	Chr12:8788257-8795789	12p13.31	Upregulated
TSPAN11	Chr12:30926428-30996599	12p11.21	Downregulated
ERBB3	Chr12:56079857-56103505	12q13.2	Upregulated
SERPINA5	Chr14:94561442-94593120	14q32.13	Upregulated
ONECUT1	Chr15:52756989-52791078	15q21.3	Downregulated
AC087632.1	Chr15:64181180-64381510	15q22.31	Downregulated
PCLAF	Chr15:64364311-64387687	15q22.31	Downregulated
ZBTB4	**Chr17:7459366-7484263**	**17p13.1**	**Downregulated**
POLR2A	**Chr17:7484366-7514616**	**17p13.1**	**Downregulated**
TNFSF12-TNFSF13	**Chr17:7549099-7561601**	**17p13.1**	**Downregulated**
TNFSF13	**Chr17:7558292-7561608**	**17p13.1**	**Downregulated**
SENP3	**Chr17:7561875-7571969**	**17p13.1**	**Downregulated**
SENP3-EIF4A1	**Chr17:7563287-7578715**	**17p13.1**	**Downregulated**
EIF4A1	**Chr17:7572706-7579005**	**17p13.1**	**Downregulated**
AC016876.2	**Chr17:7572826-7582024**	**17p13.1**	**Downregulated**
CD68	**Chr17:7579467-7582113**	**17p13.1**	**Downregulated**
AC016876.1	**Chr17:7581964-7584072**	**17p13.1**	**Downregulated**
MPDU1	**Chr17:7583529-7592789**	**17p13.1**	**Downregulated**
SOX15	**Chr17:7588178-7590170**	**17p13.1**	**Downregulated**
FXR2	**Chr17:7591230-7614871**	**17p13.1**	**Downregulated**
SAT2	**Chr17:7626234-7627876**	**17p13.1**	**Downregulated**
ATP1B2	**Chr17:7646627-7657768**	**17p13.1**	**Downregulated**
TP53	**Chr17:7661779-7687550**	**17p13.1**	**Downregulated**
WRAP53	**Chr17:7686071-7703502**	**17p13.1**	**Downregulated**
EFNB3	**Chr17:7705202-7711378**	**17p13.1**	**Downregulated**
DNAH2	**Chr17:7717354-7833744**	**17p13.1**	**Downregulated**
KDM6B	**Chr17:7839904-7854796**	**17p13.1**	**Downregulated**
NAA38	**Chr17:7856685-7885238**	**17p13.1**	**Downregulated**
CYB5D1	**Chr17:7857746-7862282**	**17p13.1**	**Downregulated**
AC104581.5	**Chr17:7858943-7866083**	**17p13.1**	**Downregulated**
PCTP	Chr17:55750979-55842830	17q22	Downregulated
TBCD	Chr17:82752064-82945922	17q25.3	Downregulated
ZNF100	Chr19:21722766-21767628	19p12	Downregulated
AC005394.2	Chr19:28435388-28727680	19q12	Downregulated
AC005524.1	Chr19:28683071-28727777	19q12	Downregulated
MX2	Chr21:41361943-41409390	21q22.3	Upregulated
SYP	ChrX:49187804-49200259	Xp11.23	Downregulated
SLITRK2	ChrX:145817832-145825842	Xq27.3	Downregulated
CLIC2	ChrX:155276211-155334657	Xq28	Downregulated
TMLHE	ChrX:155490115-155669944	Xq28	Downregulated
MT-TI	ChrMT:4263-4331	NA	Upregulated
MT-TQ	ChrMT:4329-4400	NA	Upregulated

Figure 5-4. High-confidence genes differentially expressed in *DGCR8*$^{+/-}$ and 22q11 DS neural progenitor cells (cont.)

Downregulated genes

	Exp 2 - Het1 vs WT	Exp 2 - 22q vs Cont.
Exp 1 - KOs vs WT	1.04E-42	1.33E-35
Exp 2 - Het1 vs WT	X	9.22E-117
Intersection 3 groups:	8.36E-76	

Upregulated genes

	Exp 2 - Het1 vs WT	Exp 2 - 22q vs Cont.
Exp 1 - KOs vs WT	2.75E-12	5.17E-06
Exp 2 - Het1 vs WT	X	2.42E-129
Intersection 3 groups:	1.17E-13	

Downregulated genes; without 17p13.1 genes

	Exp 2 - Het1 vs WT	Exp 2 - 22q vs Cont.
Exp 1 - KOs vs WT	4.23E-22	7.62E-19
Exp 2 - Het1 vs WT	X	6.68E-106
Intersection 3 groups:	7.71E-35	

Upregulated genes; without 17p13.1 genes

	Exp 2 - Het1 vs WT	Exp 2 - 22q vs Cont.
Exp 1 - KOs vs WT	2.75E-12	5.17E-06
Exp 2 - Het1 vs WT	X	2.42E-129
Intersection 3 groups:	1.17E-13	

Figure 5-5. Statistical significance of overlapping gene sets presented in Figure 5-4, with or without genes located within the 17p13.1 locus. P-values are presented for each overlaps, as calculated by the R package *SuperExactTest* [501]. Group tested and correspondence to Figure 5-4: *Exp 1 - KOs vs WT: EXp 1 - DGCR8 KOs vs WT (shared genes); Exp 2 - Het1 vs WT; Exp 2 - 22q vs Cont.: Exp 2 - 22q11.2 IPSCs vs Control IPSCs.*

5.5. Discussion

Differential expression analysis of DGCR8$^{+/-}$ heterozygous knock-out neural progenitor cells (NPCs) showed that the expression levels of 1744 genes were significantly altered compared to wild-type NPCs (adjusted p value $p < 0.01$). However, a principal component analysis revealed an important heterogeneity in the transcription profiles of the wild-type control lines used (Figure 5-1). Two of these wild-type lines have been derived from a similar CRISPR/Cas9 genome editing process to the DGCR8 KO lines. The first line, WTcrispr1, is a sister clone of the DGCR8 line; the second line, WTcrispr2, is from a different genome editing experiment targeting another gene. Compared to third control line, WT, they have both been submitted to different steps such as single-cell dissociation to allow colony selection, as well as antibiotic selection.

The differences in gene expression revealed by the principal components analysis could be due to off-target effects of the CRISPR/Cas9 genome editing process. The CRISPR/Cas9 editing can also potentially lead to off-target mutations due to non-specific binding of the guideRNAs used to direct the nuclease to a specific locus and induce double-strand DNA cleavage. Erroneous non-homologous end-joining repair (cf. Figure 4-2 Chapter 4 p.85) can potentially lead to insertion/deletion mutations, which can in turn affect the gene expression, protein function or introduce a premature stop-codon. This off-target effect has been shown to potentially affects many sites for each guideRNA

used, with efficiencies that can be similar than at the on-target site [459,460]. A common off-target mutation between the DGCR8 KO lines and the WTcrispr1 line generated in the same experiment could explain the clustering of these lines by principal component analysis (Figure 5-1.A). The CRISPR/Cas9 experiment conducted to knock-out DGCR8 has been performed used a modified version of the guideRNAs, by adding two extra guanine on the 5' end of the sequence [461]. It has been shown to greatly increase specificity, with off-targets insertion/deletion mutations being detected at rates comparable to deep-sequencing error rates [462]. However, due to the clustering of these cell lines, only whole-genome sequencing of the KO lines will allow to confidently discard the hypothesis that they have shared off-target mutations. Finally, all lines contain an antibiotic resistance cassette that integrated randomly in the genome rather than in the targeted locus. It is therefore possible that the random integration of the antibiotic resistance cassette could potentially affect the expression level of a gene within which it integrates, which could in turn lead to further transcriptome changes.

In order to avoid any potential bias caused by the difference in gene expression between the control lines, each of these wild-type lines were compared independently to one of the DGCR8 KO lines. This allowed to determine a list of 288 genes that are differentially expressed regardless of which control line is used. This approach reduces the power of the differential expression analysis by dividing the sample size by 3 in each analysis, however it allows greater confidence in these results that are reproducible between cell lines. Gene Ontology enrichment analysis showed that the biological process "Synapse Organisation" and "Synapse assembly" were significantly enriched for differentially expressed genes. These two terms have also been shown to be significant when testing genes that are differentially expressed in 22q11.2DS hiPSCs-derived neurons [444]. Other significant GO terms are encompassing a wide-range of biological processes, in particular when restricting to upregulated genes (Table 5-1.B). These upregulated genes might be directly relevant to DGCR8-mediated gene regulation, as deletion of DGCR8 leads to reduced microRNA levels which in turn could lead to increased microRNA targets mRNA levels. These significant GO terms include terms related to the extracellular matrix, muscle development, ossification, and metabolic processes. While the DGCR8 knock-out could affect neural differentiation and result to changes in neural-related gene expression levels, its effect on microRNA processing is likely to affect many genes in multiple biological pathways. This could explain the

diversity of significant GO terms. Enriched KEGG pathways for upregulated genes included the Hippo-signalling pathway, the TGF-beta signalling pathway, Axon guidance and ECM-receptor interaction. Interestingly, microarray-based transcriptome analysis of laser-capture microdissected cortical pyramidal neurons of schizophrenia patients (post-mortem brain tissue) found an enrichment of genes involved in TGF-beta signalling, cytoskeleton/dendritic integrity and extra-cellular matrix (ECM) regulation (as well as apoptosis and DNA damage repair) [463].

A second independent experiment was conducted to assess if the differentially expressed genes observed in neural progenitor cells differentiated from DGCR8 KO lines shared any similarity to the NPCs obtained from a 22q11.2DS hiPSC line. This revealed that about 30% of genes that were differentially expressed in the tested DGCR8 KO NPCs (916 for 3097 total) were also significantly dysregulated in the 22q11.2DS NPCs. These genes were compared to the 288 high-confidence genes differentially expressed in DGCR8 KO NPCs from the first experiment using multiple cell lines. The primary goal of this comparison was to gain more confidence in genes that might be dysregulated in 22q11.2DS NPCs, potentially due to the heterozygous deletion of DGCR8. It also allowed genes to be restricted to a smaller list of candidates that could be studied further to understand the effect of the DGCR8 deletion. This process identified a list of 63 genes. Some of these genes could be of particular interest in the context of 22q11.2 deletion syndrome. For instance, the Trimethyllysine hydroxylase epsilon (*TMLHE*) gene has been proposed as risk factor for autism, which is a symptom present in 22q11.2DS patients. However, while conducting this analysis I observed that around 40% of these genes mapped to the same chromosomal region 17p13.1. Moreover, their expression level was on average halved in DGCR8 KO NPCs or 22q11.2DS NPCs compared to their respective controls. This unexpected result could be explained by the following hypotheses:

- The presence of a copy-number variant (CNV) could have appeared “by chance”. Specifically, a hemizygous deletion at locus 17p13.1, that was then positively selected for in DGCR8 KO and 22q11DS lines. It has been shown that human embryonic in culture can accumulate genomic alterations with increased passage number, such as mitochondrial DNA mutations, promoter methylation or copy number variation [464]. In 2010, SNP arrays analysis of multiple human embryonic stem cell lines from different laboratories after different numbers of passages identified 843 CNVs with sizes ranging from 50 kb to 3 Mb [465]. The number of CNV found per sample (29 in hESCs) was similar

than found in HapMap samples (26 per samples), but there is a different distribution of the CNV with most differences occurring on chromosomes 10, 14, 20 as well as allosomes (X/Y). The most frequent regions of variants shared by the different cells lines were at the regions 1p36.13, 1p36.33, 2p11.2, 7q35, 14q32.32, 15q11.2, 21p11.2, 22q11.22, 22q11.21. Further studies in other cell lines including hiPSCs confirmed the presence of many genomic aberrations in pluripotent cell lines and that they tend to accumulate during cell culture [466]. However, in hiPSC lines it appears that there are more CNVs in early-passage cells than in intermediate passage due to a selection against cells that gain such genomic aberrations [467]. Some of the common CNVs in hESCs can be positively selected; for instance, it has been shown that the recurrent CNV at 20q11.21, first identified in 2008 [468], is present in around 25% of karyotypically normal cell lines [469]. This CNV increase cell survival due to the duplication of the *BCL2L1* gene [470].

In the present project, the deletion is at the locus 17p13.1. Deletion of this region in human stem cells has been reported in 2015 [471]. It includes the gene TP53, a known tumour suppressor gene [472]. The observed deletions at this locus were of different size but all included this gene, suggesting that its deletion probably confers a growth or survival advantage leading to positive selection. This was later confirmed that cells without a copy of the 17p13.1 or with a shRNA-mediated TP53 knock-down have a growth advantage due to increased proliferation and survival [473]. This positive selection could explain the observed presence of this deletion in NPCs derived from *DGCR8* KO lines and the 22q11.2 hiPSC line. The heterozygous loss of *DGCR8* could lead to a negative pressure on the cells, and cells randomly gaining this deletion on chromosome 17p13.1 would compensate this effect by conferring a growth advantage. This would explain why it does not seem to be present in the WTcrispr1 line that has been generated in the same experiment than the DGCR8 KO lines.

- The presence of a CNV on 17p13.1 could be due positive selection of CNVs in this region directly due to DGCR8 hemizygous deletion. While TP53 deletion could confer a growth advantage, it has been shown that other CNVs, such as the duplication at 20q11.21 [470], could also be positively selected due to increased cell survival or growth. In this project, all DGCR8 KO lines tested, as well as the 22q11.2DS line, seem to potentially present the same CNV. This could suggest that the deletion at 17p13.1 occurred preferentially when one copy of DGCR8 was deleted, compared to other CNVs. It has been shown that DROSHA, part of the microprocessor complex with DGCR8, as well as

DICER, involved in microRNA biogenesis, are both important for the DNA-damage response (DDR) of cells [474,475]. It has been proposed that a particular class of DICER- and DROSHA- dependent small non-coding RNA, called DDRNAS, help to recruits DDR factors to damaged DNA regions. Moreover, multiple microRNA have been shown to target the mRNA of genes involved in the DDR, as reviewed by Wan *et al.* [476], including TP53. Finally, TP53 is known to interact with the microprocessor complex, modulate its activity and in results affects the levels of multiple microRNAs [477]. The multiple links between the microprocessor complex, microRNAs and TP53 could suggest that in DGCR8 KO lines (and the 22q11.2DS line), cells with the deletion of TP53 (locus 17p13.1) are positively selected as a compensatory mechanism for the loss of one copy of DGCR8.

- DGCR8 deletion could lead to the epigenetic repression of a large cluster of genes located on region 17p13.1. It has been shown that large chromosomal regions including multiple genes can be repressed, notably in cancer, in a process called long-range epigenetic silencing (LRES). For instance in Wilm's tumours at the chromosomal locus 5q31 a large cluster of paralogous genes present promoter hypermetylation [478]. A study in LRES regions in prostate cancer have shown that they are also associated with histone modifications [479]. Moreover, analysis in hESCs showed that many genes within these prostate cancer LRES regions are also repressed in the embryonic stem cells. It is possible that the decrease of expression at the locus 17p13.1 is not due to a CNV but rather to an epigenetic silencing. This repression could, as mentioned above, act as a compensatory mechanism for the loss of DGCR8.

The presence of this possible artefact at chromosome 17p13.1 in all DGCR8 KO lines and 22q11.2DS lines impose caution regarding the interpretation of the enrichment analysis of genes dysregulated due to DGCR8 heterozygous knock-out. While some of the affected genes (including those within the significant overlap of differentially expressed genes between experiments after removing genes within this region (Figure 5-5)) are likely to results from the loss of one copy of DGCR8, it is not possible to determine which genes could results from this mutation and which would be due to the possible CNV at 17p13.1. Multiple hypotheses have been proposed to possibly explain the reduced expression at this locus, however further work will be needed to investigate this issue. The priority will be to assess the presence of CNVs in the tested lines, by the use of SNP arrays and/or whole-genome sequencing. To my knowledge, this CNV has

not been reported in any mouse embryonic stem cell DGCR8 KO line or in 22q11.2 DS hiPSCs lines.

6. General discussion

6.1. Summary and implication for the understanding of 22q11.2DS

This PhD thesis aimed to investigate genetic mechanisms that could explain the heterogeneity of symptoms observed in 22q11.2 deletion syndrome patients, and in particular why not all of these patients develop schizophrenia. For this I used a wide range of methods and techniques such as microarray-based genetic and epigenetic profiling, CRISPR/Cas9 genome editing and next-generation sequencing-based transcriptome analysis. The main hypothesis behind this research was the existence of genetic modifiers that have a cumulative effect with the deletion at chromosome 22q11.2DS, additively or through epistatic interactions, following a multiple-hit hypothesis. The goal of the first part of this thesis was to assess the presence of such modifiers, by looking at the burden of common polymorphisms associated with schizophrenia in 22q11.2DS patients with this neuropsychiatric disorder. It has already been suggested that copy-number variants outside of the 22q11.2 region can potentially act as a second hit in 22q11.2DS patients with schizophrenia [157,158], but the role of single-nucleotide polymorphisms had yet to be determined. The other parts of the thesis investigated possible mechanisms of interactions between the 22q11.2 deletion and genetic modifiers. The second results chapter aimed at looking at the effect of the deletion on DNA methylation and to see if epigenetic changes were affecting genes and pathways relevant to schizophrenia. The goal of the final parts of the thesis was to study the effect of *DGCR8* haploinsufficiency during cortical neuron differentiation. For this, human embryonic stem cells were altered with the CRISPR/Cas9 genome editing technology to knock-out *DGCR8* and then the generated cell lines were differentiated into neural progenitor cells. This facilitated the investigation of transcriptional changes due to the *DGCR8* mutation. *DGCR8* is a gene located within the 22q11.2 deletion and its role in genetic regulation, via the microprocessor complex that regulate microRNA biogenesis [173], makes it an interesting candidate to investigate epistatic interactions between the 22q11.2 deletion and the rest of the genome [480]. Disruption of the microRNA-buffering system can potentially lead to an uncovering of cryptic mutations, such as the common variants studied in the first result chapter.

In the first chapter, I used a polygenic risk score (PRS) analysis approach to characterise a potential role of common variants in the aetiology of schizophrenia in 22q11.2DS. I established a PRS profile of patients with or without 22q11.2DS and with or without schizophrenia to be able to compare them. It has been previously shown that

the PRS allows the discrimination of groups of individuals with or without schizophrenia in the general population [92,202], with SNPs accounting for 23% of the variation in liability in schizophrenia [201]. Results of this PhD project revealed that PRS was also allowing the discrimination of 22q11.2DS patients depending on their schizophrenia status. This indicates that common variants associated with schizophrenia do indeed play a role in the aetiology of this psychiatric disorder in patients with the 22q11.2 deletion. 22q11.2DS patients with schizophrenia have a lower PRS than schizophrenia patients without the deletion, supporting a multifactorial threshold model in which both the deletion and common polymorphisms participate to the total liability for schizophrenia.

In the second results chapter, I used DNA methylation microarrays to compare the methylome of individuals with or with the 22q11.2 deletion. It revealed that the haploinsufficiency of genes within this region leads to DNA methylation changes in other part of the genome in 22q11.2DS patients. These changes were affecting genes included in biological pathways potentially relevant to schizophrenia, such as intracellular membrane trafficking, metabolism, cell signalling and adhesion. Examples of significant pathways include the MAPK and AKT signalling pathways, known to be affected in schizophrenia [349,350]. Moreover, differentially methylated regions (regions that included multiple differentially methylated sites) have been characterised in the promoter regions of genes potentially associated with schizophrenia and/or neurodevelopment. As the DNA methylation levels of promoter regions is known to affect gene expression [248], the 22q11.2 deletion could potentially lead to an increased risk for schizophrenia through a genetic dysregulation of the expression of these genes. Furthermore, this dysregulation could interact epistatically with mutations within these genes.

In the third result chapter, I generated *DGCR8* knock-out human embryonic stem cell (hESC) lines with the CRISPR/Cas9 genome editing technology [406,407]. These cell lines exhibited molecular and cellular phenotypes, some of which have been previously demonstrated in other *DGCR8*$^{+/-}$ models. *DGCR8*$^{+/-}$ hESCs had altered levels of different levels of RNA known to be processed by DGCR8, such as microRNAs from the miR 17~92 cluster [420] but also DGCR8's own mRNA, that contain microRNA-like hairpin structures [415,481]. Differentiation of *DGCR8*$^{+/-}$ hESCs into neural progenitor cells (NPCs) suggest that their ability to process mRNA with such structures is also altered in this cell type, as shown by increased levels of *DGCR8* mRNA as well as *NEUROG2* mRNA (a pro-neuronal transcription factor whose mRNA contain hairpin structures and is regulated by the microprocessor complex).

In the final result chapter, I investigated the transcriptome of these *DGCR8*$^{+/-}$ NPCs. The haploinsufficiency of DGCR8 lead to difference of many mRNA levels, some of which are affecting pathways relevant to schizophrenia. There is also an overlap between dysregulated *DGCR8*$^{+/-}$ NPCs and NPCs differentiated from induced pluripotent stem cells derived from 22q11.2DS patients. It indicates that the dysregulation of gene expression that is observed in these patients [441] is likely due -at least in part- to the deletion of *DGCR8*. However, a potential genomic abnormality has been detected in the *DGCR8*$^{+/-}$ and 22q11.2DS cell lines used and prevents further interpretation of the results at this stage (as discussed below).

6.2. Main limitations and possible solutions

The study of 22q11.2DS-specific DNA methylation changes informs us about epigenetic changes due to the haploinsufficiency of genes within this region. They can affect genes that are associated with schizophrenia, however comparing the DNA methylation profile of patients with or without the 22q11.2 deletion does not inform about possible changes specific to patients with schizophrenia. A second analysis would be required, comparing the methylome of 22q11.2DS patients with or without this neuropsychiatric disorder. Performing such symptom-specific epigenome-wide association study of 22q11.2DS patients would help to understand the molecular mechanisms underlying the heterogeneity of clinical features observed between these patients.

The use of DNA from blood samples revealed an important bias due to the different leucocytes composition in 22q11.2DS patients (T-cell deficits). This lead to an enrichment of abnormal methylation for genes involved in immune-related biological pathways. While this can be partially accounted for by estimating the cell type composition from the methylation data and adjusting for it in the statistical analysis, it is likely that it leads to a decrease in statistical power (thus rejecting true positive differences in methylation) while keeping part of the bias (thus not rejecting all false positive differences). Although using DNA extracted from brain tissue would be more biologically relevant to study DNA methylation in the context of schizophrenia, cell composition differences between 22q11.2DS patients and controls might also occur. FACS-sorting or laser capture microdissection of a specific cell type of interest would allow the detection of differences in DNA methylation directly due to the 22q11.2 deletion without being confounded by cell type-specific effects.

Finally, a potential artefact has been detected at the locus 22q11.2. The analysis revealed a strong window of association at the 22q11.2 locus, and current data does not allow the determination of whether it is due to a technical artefact or to a true biological signal. However, re-analysis of published data from patients with or without a deletion at chromosome 7q11.23 [330] revealed a similar window of association. This indicates that such changes might be due to hemizygosity of the region, but it is not possible at this stage to determine their origin (technical or biological). The analysis of DNA methylation profiles from patients with CNVs located at other loci would help to confirm these results. Then, a more in-depth analysis of these regions should be performed using other methods that measure DNA methylation, such as targeted pyrosequencing or whole-genome reduced representation bisulfite sequencing (RRBS) [482]. It would help to understand if this artefact is of technical origin.

Note: at the time of writing this thesis, Starnawska *et al.* published a study investigating neonatal DNA methylation differences between 22q11.2DS patients that developed psychiatric disorders later on in life [483] (published online 29 August 2017). The analysis included 164 22q11.2DS patients, 48 of which have been diagnosis with a neuropsychiatric disorder. They detected significant DNA methylation changes associated with neuropsychiatric disorders in 22q11.2DS patients, including changes specific to subphenotypes such as schizophrenia or intellectual disability. They compared DNA methylation profiles between 22q11.2DS patients with the small (LCR A to LCR B) and large (LCR A to LCR D) deletions, revealing 107 differentially methylated probes, of which 100 (93%) where mapped to genes located between LCR C and LCR D (region non-overlapping between the two types of deletion). This result is particularly interesting as it replicates the effect reported in this thesis of an association of differential methylation at haploinsufficient regions, implicating that differential methylation at such loci is likely to be a technical artefact.

The transcriptome analysis of neural progenitor cells differentiated from *DGCR8*$^{+/-}$ human embryonic stem cells and 22q11.2DS human induced pluripotent stem cells revealed a region of reduced expression on chromosome 17p13.1. The decrease in expression of the 25 differentially expressed genes within this region is -2-fold, which could suggest the presence of a copy-number variation (deletion). Such a deletion at 17p13.1 has been previously reported in human embryonic stem cells [471]. However, if such a deletion occurred, it must have occurred independently in *DGCR8*$^{+/-}$ and

22q11.2DS NPCs but not in any of the control lines. Further work would be required to establish if this is coincidental or potentially indicates a genomic instability common to the *DGCR8*$^{+/-}$ and 22q11.2DS NPCs. A likely hypothesis is that the reduced expression observed at genes spanning 17p13.1 is be due to a positive selection of cells containing a deletion at this region. This could be due to pro-proliferation or pro-survival effects of the haplosuficiency of genes within it, or due to a compensation of negative effects due to the deletion of *DGCR8*. The use of genotyping microarray (or whole-genome sequencing) in the tested lines will allow the presence of a CNV to be confirmed. The next step will be to try and understand how this event occurred, and this could be achieved by characterising additional *DGCR8*$^{+/-}$ lines. Multiple heterozygous knock-out lines have been generated but not analysed (including lines not referenced in Figure 4-5.B Chapter 4 p.92 because they have not yet been sequenced for the presence of mutations on the second allele of *DGCR8*). If a CNV is confirmed, the rate of occurrence of this event should be compared between *DGCR8*$^{+/-}$ hESCs lines and sister lines generated in the same CRISPR/Cas9 editing process but without DGCR8 mutations. Similar comparisons should be done for other 22q11.2DS hiPSCs lines. This would help to understand if the CNV occurred randomly or whether it was stimulated by the presence of a *DGCR8* deletion.

6.3. Future work

The work presented in this thesis only uncovers a small part of the complex mechanisms leading to the heterogeneity of symptoms in 22q11.2DS patients. It presents the importance of the genetic background in the development of schizophrenia in these patients (common polymorphisms) as well as potential mechanisms of interactions between the deletion and the genetic background (DNA methylation changes and transcriptome changes through DGCR8-dependent microRNA dysregulation). However, an effort should be made to integrate genetic, epigenetic and transcriptomic data to try and understand the interplay between these different levels of complexity.

It has been shown that SNP alleles can be associated with different methylation levels of CpG sites [484]. These specific common polymorphisms are called methylation quantitative trait loci (mQTL). Hannon *et al.* (2016) have shown that mQTLs in human foetal brain samples are enriched for SNPs that are associated in schizophrenia in genome-wide association studies [485]. This supports the neurodevelopmental hypothesis of schizophrenia aetiology and proposes a possible epigenetic effect of genetic variation on

the development of this disorder. Moreover, this group also showed that the levels of particular CpG methylation sites were correlated with the schizophrenia polygenic risk score [486]. The schizophrenia PRS can also be associated with differential methylation in different brain regions specifically [264]. It would thus be interesting to determine if there are such associations between schizophrenia PRS and differences in methylation between 22q11.2DS patients with or without schizophrenia. Ideally, samples would be run on three types of arrays: genotyping microarrays, DNA methylation microarrays and expression microarrays (or alternatively RNA-sequencing). This would allow the investigation of the relationship between PRS and DNA methylation in 22q11.2DS, as well as the presence of mQTL in 22q11.2DS patients and determine their effect on gene expression.

In order to test the hypothesis of *DGCR8*-mediated epistatic interactions (through microRNA dysregulation), it would be interesting to investigate genetic and epigenetic variation in genes affected by *DGCR8*-dependent microRNAs specifically. Merico *et al.* have demonstrated that the genes that are predicted to be regulated by DGCR8-dependent microRNAs are significantly represented in a network of genes relevant to schizophrenia [487]. They also have identified in 22q11.2DS patients an increased burden of rare damaging variants that affect genes involved in neuronal functions; while being underpowered, the study showed that restricting the analysis to genes predicted to be impacted by a *DGCR8* deletion led to increased differences between patients with or without the deletion [160]. Restricting a polygenic risk score analysis to such genes affected by *DGCR8* haploinsufficiency might reveal an enrichment of schizophrenia-associated variation in genes targeted by microRNA in 22q11.2DS patients with schizophrenia compared to schizophrenia patients without the deletion. Similarly, analysing DNA methylation profiles for genes that are regulated by microRNAs might reveal synergistic mechanisms between DNA methylation and microRNA regulation in genes associated with schizophrenia risk.

The effect of *DGCR8* deletions on microRNA levels in human brain tissue is still unknown. Performing microRNA-sequencing on neural progenitor cells and neurons differentiated from the *DGCR8*$^{+/-}$ hESCs generated in this PhD project will help to understand how gene regulation is affected by *DGCR8* haploinsufficiency. Moreover, it will help to interpret the results of mRNA-sequencing; establishing which mRNA levels might be directly affected by abnormal microRNA levels and which mRNA levels might be changed due to more indirect effect, such as an altered neurodifferentiation.

DNA methylation sites that are the most significantly altered in 22q11.2DS patients could be specifically tested in NPC and neurons derived from 22q11.2DS hiPSCs by performing a targeted analysis (for instance with pyrosequencing). It could also be performed in *DGCR8*$^{+/-}$ NPC and neurons which would allow the investigation of a possible role of DGCR8 deletions on the epigenetic changes observed in 22q11.2DS patients.

6.4. Conclusions

The findings presented in this PhD thesis provide insights into the molecular mechanisms leading to schizophrenia and heterogeneous phenotypes in 22q11.2 deletion syndrome patients. They show the importance of genetic variation outside the deletion for the development of psychosis, and present evidence of the involvement of epigenetic changes. A *DGCR8* knock-out model have also been generated to study the changes in the regulation of gene expression in 22q11.2DS. While this model needs further characterisation, it could be used to study the role of one of the most promising 22q11.2DS candidate genes in the development of tissues affected in this syndrome (such as the brain or the heart). It could also help to understand the regulation of gene expression in a further extent by allowing the study of the canonical pathways of microRNA biogenesis in human cells. Hopefully, integration of genetic, epigenetic and transcriptomic studies will lead to a better understanding of 22q11.2DS and to the discovery of potential treatments.

7. References

1. Shprintzen, R. J. Velo-cardio-facial syndrome: 30 Years of study. *Dev. Disabil. Res. Rev.* **14,** 3–10 (2008).
2. Robin, N. H. & Shprintzen, R. J. Defining the clinical spectrum of deletion 22q11.2. *J. Pediatr.* **147,** 90–6 (2005).
3. Driscoll, D. a *et al.* Prevalence of 22q11 microdeletions in DiGeorge and velocardiofacial syndromes: implications for genetic counselling and prenatal diagnosis. *J. Med. Genet.* **30,** 813–7 (1993).
4. Goodship, J., Cross, I., LiLing, J. & Wren, C. A population study of chromosome 22q11 deletions in infancy. *Arch. Dis. Child.* **79,** 348–51 (1998).
5. Oskarsdóttir, S., Vujic, M. & Fasth, A. Incidence and prevalence of the 22q11 deletion syndrome: a population-based study in Western Sweden. *Arch. Dis. Child.* **89,** 148–51 (2004).
6. Hospital, T. *et al.* Special Article Birth Prevalence of Chromosome 22q11 . 2 Deletion Syndrome : A Systematic Review of Population-Based Studies. **99,** 187–193 (2016).
7. Maisenbacher, M. K. *et al.* Incidence of the 22q11.2 deletion in a large cohort of miscarriage samples. *Mol. Cytogenet.* **10,** 6 (2017).
8. McDonald-McGinn, D. M. *et al.* Phenotype of the 22q11.2 deletion in individuals identified through an affected relative: cast a wide FISHing net! *Genet. Med.* **3,** 23–9 (2001).
9. Digilio, M. C. *et al.* Spectrum of clinical variability in familial deletion 22q11.2: from full manifestation to extremely mild clinical anomalies. *Clin. Genet.* **63,** 308–13 (2003).
10. Thomas, N. S. *et al.* Parental and chromosomal origins of microdeletion and duplication syndromes involving 7q11.23, 15q11-q13 and 22q11. *Eur. J. Hum. Genet.* **14,** 831–7 (2006).
11. Delio, M. *et al.* Enhanced maternal origin of the 22q11.2 deletion in velocardiofacial and DiGeorge syndromes. *Am. J. Hum. Genet.* **92,** 439–47 (2013).
12. Carlson, C. *et al.* Molecular definition of 22q11 deletions in 151 velo-cardio-facial syndrome patients. *Am. J. Hum. Genet.* **61,** 620–9 (1997).
13. Lindsay, E. A. *et al.* Velo-cardio-facial syndrome: frequency and extent of 22q11 deletions. *Am. J. Med. Genet.* **57,** 514–22 (1995).
14. Lindsay, E. A. *et al.* Submicroscopic deletions at 22q11.2: variability of the clinical picture and delineation of a commonly deleted region. *Am. J. Med. Genet.* **56,** 191–7 (1995).
15. Shaikh, T. H. *et al.* Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* **9,** 489–501 (2000).
16. Kurahashi, H. *et al.* Another critical region for deletion of 22q11: a study of 100 patients. *Am. J. Med. Genet.* **72,** 180–5 (1997).
17. Burnside, R. D. 22q11.21 Deletion Syndromes: A Review of Proximal, Central, and Distal Deletions and Their Associated Features. *Cytogenet. Genome Res.* **146,** 89–99 (2015).
18. Edelmann, L., Pandita, R. K. & Morrow, B. E. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am. J. Hum. Genet.* **64,** 1076–86 (1999).
19. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1,** 4 (2008).
20. Saitta, S. C. *et al.* Aberrant interchromosomal exchanges are the predominant cause of the 22q11.2 deletion. *Hum. Mol. Genet.* **13,** 417–28 (2004).
21. Ben-Shachar, S. *et al.* 22q11.2 distal deletion: a recurrent genomic disorder distinct from DiGeorge syndrome and velocardiofacial syndrome. *Am. J. Hum. Genet.* **82,** 214–21 (2008).
22. Mikhail, F. M. *et al.* The recurrent distal 22q11.2 microdeletions are often de novo and

do not represent a single clinical entity: a proposed categorization system. *Genet. Med.* **16,** 92–100 (2014).
23. Ensenauer, R. E. *et al.* Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. *Am. J. Hum. Genet.* **73,** 1027–40 (2003).
24. Torres-Juan, L., Rosell, J., Sánchez-de-la-Torre, M., Fibla, J. & Heine-Suñer, D. Analysis of meiotic recombination in 22q11.2, a region that frequently undergoes deletions and duplications. *BMC Med. Genet.* **8,** 14 (2007).
25. Adeyinka, A. *et al.* Familial 22q11.2 deletions in DiGeorge/velocardiofacial syndrome are predominantly smaller than the commonly observed 3Mb. *Genet. Med.* **6,** 517–20 (2004).
26. Lindsay, E. A., Halford, S., Wadey, R., Scambler, P. J. & Baldini, A. Molecular cytogenetic characterization of the DiGeorge syndrome region using fluorescence in situ hybridization. *Genomics* **17,** 403–7 (1993).
27. Fernández, L. *et al.* Comparative study of three diagnostic approaches (FISH, STRs and MLPA) in 30 patients with 22q11.2 deletion syndrome. *Clin. Genet.* **68,** 373–378 (2005).
28. Jalali, G. R. *et al.* Detailed analysis of 22q11.2 with a high density MLPA probe set. *Hum. Mutat.* **29,** 433–40 (2008).
29. Sørensen, K. M. *et al.* Detecting 22q11.2 deletions by use of multiplex ligation-dependent probe amplification on DNA from neonatal dried blood spot samples. *J. Mol. Diagn.* **12,** 147–51 (2010).
30. Weksberg, R. *et al.* A method for accurate detection of genomic microdeletions using real-time quantitative PCR. *BMC Genomics* **6,** 180 (2005).
31. Tomita-Mitchell, A. *et al.* Multiplexed quantitative real-time PCR to detect 22q11.2 deletion in patients with congenital heart disease. *Physiol. Genomics* **42A,** 52–60 (2010).
32. Bittel, D. C. *et al.* Refining the 22q11.2 deletion breakpoints in DiGeorge syndrome by aCGH. *Cytogenet. Genome Res.* **124,** 113–20 (2009).
33. Jensen, T. J., Dzakula, Z., Deciu, C., van den Boom, D. & Ehrich, M. Detection of microdeletion 22q11.2 in a fetus by next-generation sequencing of maternal plasma. *Clin. Chem.* **58,** 1148–51 (2012).
34. Wilson, D. I., Burn, J., Scambler, P. & Goodship, J. DiGeorge syndrome: part of CATCH 22. *J. Med. Genet.* **30,** 852–6 (1993).
35. Momma, K. Cardiovascular anomalies associated with chromosome 22q11.2 deletion syndrome. *Am. J. Cardiol.* **105,** 1617–24 (2010).
36. Momma, K. Cardiovascular anomalies associated with chromosome 22q11.2 deletion. *Int. J. Cardiol.* **114,** 147–9 (2007).
37. Lee, M.-Y. *et al.* Variety of prenatally diagnosed congenital heart disease in 22q11.2 deletion syndrome. *Obstet. Gynecol. Sci.* **57,** 11–6 (2014).
38. Óskarsdóttir, S., Holmberg, E., Fasth, A. & Strömland, K. Facial features in children with the 22q11 deletion syndrome. *Acta Paediatr. Int. J. Paediatr.* **97,** 1113–1117 (2008).
39. Becker, D. B. *et al.* Accuracy in identification of patients with 22q11.2 deletion by likely care providers using facial photographs. *Plast. Reconstr. Surg.* **114,** 1367–1372 (2004).
40. McDonald-McGinn, D. M. *et al.* The Philadelphia story: the 22q11.2 deletion: report on 250 patients. *Genet. Couns.* **10,** 11–24 (1999).
41. Ryan, a K. *et al.* Spectrum of clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study. *J. Med. Genet.* **34,** 798–804 (1997).
42. Pasick, C. *et al.* Asymmetric crying facies in the 22q11.2 deletion syndrome: implications for future screening. *Clin. Pediatr. (Phila).* **52,** 1144–8 (2013).
43. Jawad, A. F., McDonald-Mcginn, D. M., Zackai, E. & Sullivan, K. E. Immunologic features of chromosome 22q11.2 deletion syndrome (DiGeorge syndrome/velocardiofacial syndrome). *J. Pediatr.* **139,** 715–23 (2001).
44. Piliero, L. M., Sanford, A. N., McDonald-McGinn, D. M., Zackai, E. H. & Sullivan, K. E. T-cell homeostasis in humans with thymic hypoplasia due to chromosome 22q11.2

deletion syndrome. *Blood* **103,** 1020–5 (2004).
45. McLean-Tooke, A., Spickett, G. P. & Gennery, A. R. Immunodeficiency and autoimmunity in 22q11.2 deletion syndrome. *Scand. J. Immunol.* **66,** 1–7 (2007).
46. Lima, K. *et al.* Hypoparathyroidism and autoimmunity in the 22q11.2 deletion syndrome. *Eur. J. Endocrinol.* **165,** 345–352 (2011).
47. Cheung, E. N. M. *et al.* Prevalence of hypocalcaemia and its associated features in 22q11·2 deletion syndrome. *Clin. Endocrinol. (Oxf).* **81,** 190–196 (2014).
48. Brauner, R. *et al.* Parathyroid function and growth in 22q11.2 deletion syndrome. *J. Pediatr.* **142,** 504–508 (2003).
49. Weinzimer, S. A. *et al.* Growth hormone deficiency in patients with 22q11.2 deletion: expanding the phenotype. *Pediatrics* **101,** 929–932 (1998).
50. Tarquinio, D. C., Jones, M. C., Jones, K. L. & Bird, L. M. Growth charts for 22q11 deletion syndrome. *Am. J. Med. Genet. Part A* **158 A,** 2672–2681 (2012).
51. Shprintzen, R. J. *et al.* A new syndrome involving cleft palate, cardiac anomalies, typical facies, and learning disabilities: velo-cardio-facial syndrome. *Cleft Palate J.* **15,** 56–62 (1978).
52. Niarchou, M. *et al.* Psychopathology and cognition in children with 22q11.2 deletion syndrome. *Br. J. Psychiatry* **204,** 46–54 (2014).
53. Niklasson, L., Rasmussen, P., Oskarsdóttir, S. & Gillberg, C. Neuropsychiatric disorders in the 22q11 deletion syndrome. *Genet. Med.* **3,** 79–84 (2001).
54. Fung, W. L. A. *et al.* Elevated prevalence of generalized anxiety disorder in adults with 22q11.2 deletion syndrome. *Am. J. Psychiatry* **167,** 998 (2010).
55. Jonas, R. K., Montojo, C. A. & Bearden, C. E. The 22q11.2 deletion syndrome as a window into complex neuropsychiatric disorders over the lifespan. *Biol. Psychiatry* **75,** 351–60 (2014).
56. Philip, N. & Bassett, A. Cognitive, behavioural and psychiatric phenotype in 22q11.2 deletion syndrome. *Behav. Genet.* **41,** 403–12 (2011).
57. Schneider, M. *et al.* Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: results from the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. *Am. J. Psychiatry* **171,** 627–639 (2014).
58. Bassett, A. S. *et al.* Clinical features of 78 adults with 22q11 Deletion Syndrome. *Am. J. Med. Genet. A* **138,** 307–13 (2005).
59. Murphy, K. C., Jones, L. a & Owen, M. J. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch. Gen. Psychiatry* **56,** 940–5 (1999).
60. Moss, E. M. *et al.* Psychoeducational profile of the 22q11.2 microdeletion: A complex pattern. *J. Pediatr.* **134,** 193–8 (1999).
61. Vorstman, J. A. S. *et al.* Cognitive decline preceding the onset of psychosis in patients with 22q11.2 deletion syndrome. *JAMA psychiatry* **72,** 377–85 (2015).
62. Chawner, S. J. R. A. *et al.* Childhood cognitive development in 22q11.2 deletion syndrome: case-control study. *Br. J. Psychiatry* bjp.bp.116.195651 (2017). doi:10.1192/bjp.bp.116.195651
63. Zaleski, C. *et al.* The co-occurrence of early onset Parkinson disease and 22q11.2 deletion syndrome. *Am. J. Med. Genet. A* **149A,** 525–8 (2009).
64. Rehman, A. F., Dhamija, R., Williams, E. S. & Barrett, M. J. 22q11.2 deletion syndrome presenting with early-onset Parkinson's disease. *Mov. Disord.* **30,** 1289–90 (2015).
65. Oki, M. *et al.* Early-onset Parkinson's Disease Associated with Chromosome 22q11.2 Deletion Syndrome. *Intern. Med.* **55,** 303–5 (2016).
66. Dufournet, B. *et al.* Parkinson's disease associated with 22q11.2 deletion: Clinical characteristics and response to treatment. *Rev. Neurol. (Paris).* **173,** 406–410 (2017).
67. Butcher, N. J. *et al.* Association between early-onset Parkinson disease and 22q11.2 deletion syndrome: identification of a novel genetic form of Parkinson disease and its clinical implications. *JAMA Neurol.* **70,** 1359–66 (2013).
68. Mok, K. Y. *et al.* Deletions at 22q11.2 in idiopathic Parkinson's disease: a combined analysis of genome-wide association data. *Lancet. Neurol.* **15,** 585–96 (2016).

69. Repetto, G. M. *et al.* Case fatality rate and associated factors in patients with 22q11 microdeletion syndrome: a retrospective cohort study. *BMJ Open* **4,** e005041 (2014).
70. Bassett, A. S. *et al.* Premature death in adults with 22q11.2 deletion syndrome. *J. Med. Genet.* **46,** 324–30 (2009).
71. Michaelovsky, E. *et al.* Genotype-phenotype correlation in 22q11.2 deletion syndrome. *BMC Med. Genet.* **13,** 122 (2012).
72. Sandrin-Garcia, P. *et al.* Typical phenotypic spectrum of velocardiofacial syndrome occurs independently of deletion size in chromosome 22q11.2. *Mol. Cell. Biochem.* **303,** 9–17 (2007).
73. Rauch, A. *et al.* Systematic assessment of atypical deletions reveals genotype-phenotype correlation in 22q11.2. *J. Med. Genet.* **42,** 871–6 (2005).
74. Guna, A., Butcher, N. J. & Bassett, A. S. Comparative mapping of the 22q11.2 deletion region and the potential of simple model organisms. *J. Neurodev. Disord.* **7,** 18 (2015).
75. Lindsay, E. a *et al.* Congenital heart disease in mice deficient for the DiGeorge syndrome region. *Nature* **401,** 379–83 (1999).
76. Merscher, S. *et al.* TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104,** 619–29 (2001).
77. Paylor, R. *et al.* Tbx1 haploinsufficiency is linked to behavioral disorders in mice and humans: implications for 22q11 deletion syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 7729–34 (2006).
78. Stark, K. L. *et al.* Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat. Genet.* **40,** 751–60 (2008).
79. Swerdlow, N. R., Weber, M., Qu, Y., Light, G. A. & Braff, D. L. Realistic expectations of prepulse inhibition in translational models for schizophrenia research. *Psychopharmacology (Berl).* **199,** 331–88 (2008).
80. Sobin, C., Kiley-Brabeck, K. & Karayiorgou, M. Lower prepulse inhibition in children with the 22q11 deletion syndrome. *Am. J. Psychiatry* **162,** 1090–9 (2005).
81. Earls, L. R. *et al.* Dysregulation of presynaptic calcium and synaptic plasticity in a mouse model of 22q11 deletion syndrome. *J. Neurosci.* **30,** 15843–55 (2010).
82. Drew, L. J. *et al.* Evidence for altered hippocampal function in a mouse model of the human 22q11.2 microdeletion. *Mol. Cell. Neurosci.* **47,** 293–305 (2011).
83. Earls, L. R. *et al.* Age-dependent microRNA control of synaptic plasticity in 22q11 deletion syndrome and schizophrenia. *J. Neurosci.* **32,** 14132–44 (2012).
84. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30,** 67–76 (2008).
85. Owen, M. J. & Doherty, J. L. What can we learn from the high rates of schizophrenia in people with 22q11.2 deletion syndrome? *World Psychiatry* **15,** 23–5 (2016).
86. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204,** 108–14 (2014).
87. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49,** 27–35 (2016).
88. Männistö, P. T. & Kaakkola, S. Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors. *Pharmacol. Rev.* **51,** 593–628 (1999).
89. Lachman, H. M. *et al.* Human catechol-O-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. *Pharmacogenetics* **6,** 243–50 (1996).
90. Chen, J. *et al.* Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. *Am. J. Hum. Genet.* **75,** 807–21 (2004).
91. Williams, H. J., Owen, M. J. & O'Donovan, M. C. Is COMT a susceptibility gene for schizophrenia? *Schizophr. Bull.* **33,** 635–41 (2007).
92. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–7 (2014).

93. González-Castro, T. B. *et al.* The Role of a Catechol-O-Methyltransferase (COMT) Val158Met Genetic Polymorphism in Schizophrenia: A Systematic Review and Updated Meta-analysis on 32,816 Subjects. *Neuromolecular Med.* **18,** 216–31 (2016).
94. Nohesara, S. *et al.* DNA hypomethylation of MB-COMT promoter in the DNA derived from saliva in schizophrenia and bipolar disorder. *J. Psychiatr. Res.* **45,** 1432–8 (2011).
95. Abdolmaleky, H. M. *et al.* Hypomethylation of MB-COMT promoter is a major risk factor for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* **15,** 3132–45 (2006).
96. Melas, P. A. *et al.* Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. *FASEB J.* **26,** 2712–8 (2012).
97. Gogos, J. A. *et al.* The gene encoding proline dehydrogenase modulates sensorimotor gating in mice. *Nat. Genet.* **21,** 434–9 (1999).
98. Jacquet, H. *et al.* PRODH mutations and hyperprolinemia in a subset of schizophrenic patients. *Hum. Mol. Genet.* **11,** 2243–9 (2002).
99. Raux, G. *et al.* Involvement of hyperprolinemia in cognitive and psychiatric features of the 22q11 deletion syndrome. *Hum. Mol. Genet.* **16,** 83–91 (2007).
100. Willis, A., Bender, H. U., Steel, G. & Valle, D. PRODH variants and risk for schizophrenia. *Amino Acids* **35,** 673–9 (2008).
101. Papaioannou, V. E. The T-box gene family: emerging roles in development, stem cells and cancer. *Development* **141,** 3819–33 (2014).
102. Lindsay, E. a *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410,** 97–101 (2001).
103. Jerome, L. A. & Papaioannou, V. E. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat. Genet.* **27,** 286–91 (2001).
104. Vitelli, F., Morishima, M., Taddei, I., Lindsay, E. A. & Baldini, A. Tbx1 mutation causes multiple cardiovascular defects and disrupts neural crest and cranial nerve migratory pathways. *Hum. Mol. Genet.* **11,** 915–22 (2002).
105. Foster, K. *et al.* Contribution of neural crest-derived cells in the embryonic and adult thymus. *J. Immunol.* **180,** 3183–9 (2008).
106. Keyte, A. & Hutson, M. R. The neural crest in cardiac congenital anomalies. *Differentiation.* **84,** 25–40 (2012).
107. Cordero, D. R. *et al.* Cranial neural crest cells on the move: their roles in craniofacial development. *Am. J. Med. Genet. A* **155A,** 270–9 (2011).
108. Hiramoto, T. *et al.* Tbx1: identification of a 22q11.2 gene as a risk factor for autism spectrum disorder in a mouse model. *Hum. Mol. Genet.* **20,** 4775–85 (2011).
109. Flore, G., Cioffi, S., Bilio, M. & Illingworth, E. Cortical Development Requires Mesodermal Expression of Tbx1, a Gene Haploinsufficient in 22q11.2 Deletion Syndrome. *Cereb. Cortex* **27,** 2210–2225 (2017).
110. Zweier, C., Sticht, H., Aydin-Yaylagül, I., Campbell, C. E. & Rauch, A. Human TBX1 missense mutations cause gain of function resulting in the same phenotype as 22q11.2 deletions. *Am. J. Hum. Genet.* **80,** 510–7 (2007).
111. Torres-Juan, L. *et al.* Mutations in TBX1 genocopy the 22q11.2 deletion and duplication syndromes: a new susceptibility factor for mental retardation. *Eur. J. Hum. Genet.* **15,** 658–63 (2007).
112. Li, Y. *et al.* Association study between GNB1L and three major mental disorders in Chinese Han populations. *Psychiatry Res.* **187,** 457–9 (2011).
113. Williams, N. M. *et al.* Strong evidence that GNB1L is associated with schizophrenia. *Hum. Mol. Genet.* **17,** 555–66 (2008).
114. Sun, Y., Tao, Y., Wang, J. & Saffen, D. The schizophrenia/bipolar disorder candidate gene GNB1L is regulated in human temporal cortex by a cis-acting element located within the 3'-region. *Neurosci. Bull.* **31,** 43–52 (2015).
115. Ishiguro, H. *et al.* Supportive evidence for reduced expression of GNB1L in schizophrenia. *Schizophr. Bull.* **36,** 756–65 (2010).
116. Liu, H. *et al.* Genetic variation in the 22q11 locus and susceptibility to schizophrenia.

Proc. Natl. Acad. Sci. U. S. A. **99,** 16859–64 (2002).
117. Chen, W.-Y. *et al.* Case-control study and transmission disequilibrium test provide consistent evidence for association between schizophrenia and genetic variation in the 22q11 gene ZDHHC8. *Hum. Mol. Genet.* **13,** 2991–5 (2004).
118. Mukai, J. *et al.* Evidence that the gene encoding ZDHHC8 contributes to the risk of schizophrenia. *Nat. Genet.* **36,** 725–31 (2004).
119. Glaser, B. *et al.* No association between the putative functional ZDHHC8 single nucleotide polymorphism rs175174 and schizophrenia in large European samples. *Biol. Psychiatry* **58,** 78–80 (2005).
120. Fukata, Y. & Fukata, M. Protein palmitoylation in neuronal development and synaptic plasticity. *Nat. Rev. Neurosci.* **11,** 161–75 (2010).
121. Mukai, J. *et al.* Molecular substrates of altered axonal growth and brain connectivity in a mouse model of schizophrenia. *Neuron* **86,** 680–95 (2015).
122. Mukai, J. *et al.* Palmitoylation-dependent neurodevelopmental deficits in a mouse model of 22q11 microdeletion. *Nat. Neurosci.* **11,** 1302–10 (2008).
123. Moutin, E. *et al.* Palmitoylation of cdc42 Promotes Spine Stabilization and Rescues Spine Density Deficit in a Mouse Model of 22q11.2 Deletion Syndrome. *Cereb. Cortex* 1–12 (2016). doi:10.1093/cercor/bhw183
124. Shifman, S. *et al.* A complete genetic association scan of the 22q11 deletion region and functional evidence reveal an association between DGCR2 and schizophrenia. *Hum. Genet.* **120,** 160–70 (2006).
125. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* **43,** 864–8 (2011).
126. Han, J. *et al.* The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* **18,** 3016–27 (2004).
127. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **5,** 522–31 (2004).
128. Wang, Y., Medvid, R., Melton, C., Jaenisch, R. & Blelloch, R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat. Genet.* **39,** 380–5 (2007).
129. Sellier, C. *et al.* Decreased DGCR8 expression and miRNA dysregulation in individuals with 22q11.2 deletion syndrome. *PLoS One* **9,** e103884 (2014).
130. Smalheiser, N. R. *et al.* Expression of microRNAs and other small RNAs in prefrontal cortex in schizophrenia, bipolar disorder and depressed subjects. *PLoS One* **9,** e86469 (2014).
131. Moreau, M. P., Bruse, S. E., David-Rus, R., Buyske, S. & Brzustowicz, L. M. Altered microRNA expression profiles in postmortem brain samples from individuals with schizophrenia and bipolar disorder. *Biol. Psychiatry* **69,** 188–93 (2011).
132. Kim, A. H. *et al.* MicroRNA expression profiling in the prefrontal cortex of individuals affected with schizophrenia and bipolar disorders. *Schizophr. Res.* **124,** 183–91 (2010).
133. Perkins, D. O. *et al.* microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol.* **8,** R27 (2007).
134. Beveridge, N. J. & Cairns, M. J. MicroRNA dysregulation in schizophrenia. *Neurobiol. Dis.* **46,** 263–71 (2012).
135. Lopez-Rivera, E. *et al.* Genetic Drivers of Kidney Defects in the DiGeorge Syndrome. *N. Engl. J. Med.* **376,** 742–754 (2017).
136. Weksberg, R. *et al.* Molecular characterization of deletion breakpoints in adults with 22q11 deletion syndrome. *Hum. Genet.* **120,** 837–45 (2007).
137. Bertini, V. *et al.* Deletion Extents Are Not the Cause of Clinical Variability in 22q11.2 Deletion Syndrome: Does the Interaction between DGCR8 and miRNA-CNVs Play a Major Role? *Front. Genet.* **8,** 47 (2017).
138. Rice, A. M. & McLysaght, A. Dosage-sensitive genes in evolution and disease. *BMC Biol.* **15,** 78 (2017).
139. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol.*

Psychiatry **19,** 37–40 (2014).
140. Skelly, D. A., Ronald, J. & Akey, J. M. Inherited variation in gene expression. *Annu. Rev. Genomics Hum. Genet.* **10,** 313–32 (2009).
141. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33 Suppl,** 245–54 (2003).
142. NORDLING, C. O. A new theory on cancer-inducing mechanism. *Br. J. Cancer* **7,** 68–72 (1953).
143. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68,** 820–3 (1971).
144. Bayer, T. A., Falkai, P. & Maier, W. Genetic and non-genetic vulnerability factors in schizophrenia: the basis of the 'two hit hypothesis'. *J. Psychiatr. Res.* **33,** 543–8 (1999).
145. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42,** 203–9 (2010).
146. Hoogendoorn, B. *et al.* Functional analysis of polymorphisms in the promoter regions of genes on 22q11. *Hum. Mutat.* **24,** 35–42 (2004).
147. Hestand, M. S. *et al.* A catalog of hemizygous variation in 127 22q11 deletion patients. *Hum. genome Var.* **3,** 15065 (2016).
148. Guipponi, M. *et al.* No evidence for the presence of genetic variants predisposing to psychotic disorders on the non-deleted 22q11.2 allele of VCFS patients. *Transl. Psychiatry* **7,** e1039 (2017).
149. Gothelf, D. *et al.* Association of the low-activity COMT 158Met allele with ADHD and OCD in subjects with velocardiofacial syndrome. *Int. J. Neuropsychopharmacol.* **10,** 301–8 (2007).
150. Bearden, C. E. *et al.* Effects of a functional COMT polymorphism on prefrontal cognitive function in patients with 22q11.2 deletion syndrome. *Am. J. Psychiatry* **161,** 1700–2 (2004).
151. Baker, K., Baldeweg, T., Sivagnanasundaram, S., Scambler, P. & Skuse, D. COMT Val108/158 Met modifies mismatch negativity and cognitive function in 22q11 deletion syndrome. *Biol. Psychiatry* **58,** 23–31 (2005).
152. Shashi, V. *et al.* Cognitive correlates of a functional COMT polymorphism in children with 22q11.2 deletion syndrome. *Clin. Genet.* **69,** 234–8 (2006).
153. Franconi, C. P. *et al.* IQ and hemizygosity for the Val(158) Met functional polymorphism of COMT in 22q11DS. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **171,** 1112–1115 (2016).
154. Rees, E., O'Donovan, M. C. & Owen, M. J. Genetics of schizophrenia. *Curr. Opin. Behav. Sci.* **2,** 8–14 (2015).
155. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16,** 172–83 (2015).
156. Kirov, G. CNVs in neuropsychiatric disorders. *Hum. Mol. Genet.* **24,** R45-9 (2015).
157. Williams, H. J. *et al.* Schizophrenia two-hit hypothesis in velo-cardio facial syndrome. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **162,** 177–182 (2013).
158. Bassett, A. S. *et al.* Rare Genome-Wide Copy Number Variation and Expression of Schizophrenia in 22q11.2 Deletion Syndrome. *Am. J. Psychiatry* appiajp201716121417 (2017). doi:10.1176/appi.ajp.2017.16121417
159. Mlynarski, E. E. *et al.* Copy-Number Variation of the Glucose Transporter Gene SLC2A3 and Congenital Heart Defects in the 22q11.2 Deletion Syndrome. *Am. J. Hum. Genet.* **96,** 753–64 (2015).
160. Merico, D. *et al.* Whole-Genome Sequencing Suggests Schizophrenia Risk Mechanisms in Humans with 22q11.2 Deletion Syndrome. *G3 (Bethesda).* **5,** 2453–61 (2015).
161. Tansey, K. E. *et al.* Common alleles contribute to schizophrenia in CNV carriers. *Mol. Psychiatry* **21,** 1085–9 (2016).
162. Goodship, J., Cross, I., Scambler, P. & Burn, J. Monozygotic twins with chromosome 22q11 deletion and discordant phenotype. *J. Med. Genet.* **32,** 746–8 (1995).
163. Yamagishi, H. *et al.* Phenotypic discordance in monozygotic twins with 22q11.2

deletion. *Am. J. Med. Genet.* **78,** 319–21 (1998).
164. Hillebrand, G., Siebert, R., Simeoni, E. & Santer, R. DiGeorge syndrome with discordant phenotype in monozygotic twins. *J. Med. Genet.* **37,** E23 (2000).
165. Halder, A., Jain, M., Chaudhary, I. & Varma, B. Chromosome 22q11.2 microdeletion in monozygotic twins with discordant phenotype and deletion size. *Mol. Cytogenet.* **5,** 13 (2012).
166. Brown, A. S. The environment and susceptibility to schizophrenia. *Prog. Neurobiol.* **93,** 23–58 (2011).
167. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* **13,** 97–109 (2012).
168. Phillips, P. C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9,** 855–67 (2008).
169. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15,** 722–33 (2014).
170. Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27,** 323–331 (2011).
171. Paterlini, M. *et al.* Transcriptional and behavioral interaction between 22q11.2 orthologs modulates schizophrenia-related phenotypes in mice. *Nat. Neurosci.* **8,** 1586–94 (2005).
172. Hidding, E., Swaab, H., de Sonneville, L. M. J., van Engeland, H. & Vorstman, J. A. S. The role of COMT and plasma proline in the variable penetrance of autistic spectrum symptoms in 22q11.2 deletion syndrome. *Clin. Genet.* **90,** 420–427 (2016).
173. Gregory, R. I. *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432,** 235–40 (2004).
174. Ebert, M. S. & Sharp, P. a. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149,** 515–24 (2012).
175. Grice, S. J., Liu, J.-L. & Webber, C. Synergistic Interactions between Drosophila Orthologues of Genes Spanned by De Novo Human CNVs Support Multiple-Hit Models of Autism. *PLOS Genet.* **11,** e1004998 (2015).
176. Xu, B., Hsu, P.-K., Stark, K. L., Karayiorgou, M. & Gogos, J. a. Derepression of a neuronal inhibitor due to miRNA dysregulation in a schizophrenia-related microdeletion. *Cell* **152,** 262–75 (2013).
177. Qadir, X. V, Han, C., Lu, D., Zhang, J. & Wu, T. miR-185 inhibits hepatocellular carcinoma growth by targeting the DNMT1/PTEN/Akt pathway. *Am. J. Pathol.* **184,** 2355–64 (2014).
178. Zhang, Z. *et al.* MiR-185 targets the DNA methyltransferases 1 and regulates global DNA methylation in human glioma. *Mol. Cancer* **10,** 124 (2011).
179. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2,** e00523 (2013).
180. Murray, R. M. & Lewis, S. W. Is schizophrenia a neurodevelopmental disorder? *BMJ* **295,** 681–682 (1987).
181. Rapoport, J. L., Giedd, J. N. & Gogtay, N. Neurodevelopmental model of schizophrenia: update 2012. *Mol. Psychiatry* **17,** 1228–38 (2012).
182. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–72 (2007).
183. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–7 (1998).
184. Musunuru, K. Genome editing of human pluripotent stem cells to generate human cellular disease models. *Dis. Model. Mech.* **6,** 896–904 (2013).
185. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27,** 275–80 (2009).
186. Mertens, J., Marchetto, M. C., Bardy, C. & Gage, F. H. Evaluating cell reprogramming, differentiation and conversion technologies in neuroscience. *Nat. Rev. Neurosci.* **17,** 424–37 (2016).

187. Marchetto, M. C. N. *et al.* A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143,** 527–39 (2010).
188. Wen, Z. *et al.* Synaptic dysregulation in a human iPS cell model of mental disorders. *Nature* **515,** 414–8 (2014).
189. Srikanth, P. & Young-Pearse, T. L. Stem cells on the brain: modeling neurodevelopmental and neurodegenerative diseases using human induced pluripotent stem cells. *J. Neurogenet.* **28,** 5–29 (2014).
190. Srikanth, P. *et al.* Genomic DISC1 Disruption in hiPSCs Alters Wnt Signaling and Neural Cell Fate. *Cell Rep.* **12,** 1414–29 (2015).
191. Topol, A. *et al.* Dysregulation of miRNA-9 in a Subset of Schizophrenia Patient-Derived Neural Progenitor Cells. *Cell Rep.* **15,** 1024–1036 (2016).
192. Brennand, K. *et al.* Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. *Mol. Psychiatry* **20,** 361–8 (2015).
193. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17,** 1520–8 (2007).
194. Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry.* **55,** 1068–87 (2014).
195. Cardno, A. G. & Gottesman, I. I. Twin studies of schizophrenia: From bow-and-arrow concordances to star wars Mx and functional genomics. *Am. J. Med. Genet. - Semin. Med. Genet.* **97,** 12–17 (2000).
196. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60,** 1187–92 (2003).
197. O'Rourke, D. H., Gottesman, I. I., Suarez, B. K., Rice, J. & Reich, T. Refutation of the general single-locus model for the etiology of schizophrenia. *Am. J. Hum. Genet.* **34,** 630–49 (1982).
198. Gottesman, I. . I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 199–205 (1967).
199. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8,** e1002822 (2012).
200. Fraser, F. C. The multifactorial/threshold concept -- uses and misuses. *Teratology* **14,** 267–80 (1976).
201. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44,** 247–50 (2012).
202. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–52 (2009).
203. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The lancet. Psychiatry* **4,** 57–62 (2017).
204. Hubbard, L. *et al.* Evidence of Common Genetic Overlap Between Schizophrenia and Cognition. *Schizophr. Bull.* **42,** 832–42 (2016).
205. Jones, H. J. *et al.* Phenotypic Manifestation of Genetic Risk for Schizophrenia During Adolescence in the General Population. *JAMA psychiatry* **73,** 221–8 (2016).
206. Ahn, K., An, S. S., Shugart, Y. Y. & Rapoport, J. L. Common polygenic variation and risk for childhood-onset schizophrenia. *Mol. Psychiatry* **21,** 94–6 (2016).
207. Vassos, E. *et al.* An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biol. Psychiatry* **81,** 470–477 (2017).
208. Bigdeli, T. B. *et al.* Genome-wide association study reveals greater polygenic loading for schizophrenia in cases with a family history of illness. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **171B,** 276–89 (2016).
209. Terwisscha van Scheltinga, A. F. *et al.* Genetic schizophrenia risk variants jointly modulate total brain and white matter volume. *Biol. Psychiatry* **73,** 525–31 (2013).
210. Oertel-Knöchel, V. *et al.* Schizophrenia risk variants modulate white matter volume across the psychosis spectrum: evidence from two independent cohorts. *NeuroImage. Clin.* **7,** 764–70 (2015).
211. Caseras, X., Tansey, K. E., Foley, S. & Linden, D. Association between genetic risk scoring

for schizophrenia and bipolar disorder with regional subcortical volumes. *Transl. Psychiatry* **5,** e692 (2015).
212. Papiol, S. *et al.* Polygenic determinants of white matter volume derived from GWAS lack reproducibility in a replicate sample. *Transl. Psychiatry* **4,** e362 (2014).
213. Van der Auwera, S. *et al.* Predicting brain structure in population-based samples with biologically informed genetic scores for schizophrenia. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **174,** 324–332 (2017).
214. Van der Auwera, S. *et al.* No association between polygenic risk for schizophrenia and brain volume in the general population. *Biol. Psychiatry* **78,** e41-2 (2015).
215. Harrisberger, F. *et al.* Impact of polygenic schizophrenia-related risk and hippocampal volumes on the onset of psychosis. *Transl. Psychiatry* **6,** e868 (2016).
216. Lancaster, T. M. *et al.* Associations between polygenic risk for schizophrenia and brain function during probabilistic learning in healthy individuals. *Hum. Brain Mapp.* **37,** 491–500 (2016).
217. Walton, E. *et al.* Prefrontal inefficiency is associated with polygenic risk for schizophrenia. *Schizophr. Bull.* **40,** 1263–71 (2014).
218. Hamshere, M. L. *et al.* Polygenic dissection of the bipolar phenotype. *Br. J. Psychiatry* **198,** 284–8 (2011).
219. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)* **381,** 1371–9 (2013).
220. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19,** 1017–1024 (2014).
221. Costas, J. *et al.* Exon-focused genome-wide association study of obsessive-compulsive disorder and shared polygenic risk with schizophrenia. *Transl. Psychiatry* **6,** e768 (2016).
222. Hamshere, M. L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol. Psychiatry* **18,** 708–12 (2013).
223. Rees, E. *et al.* Analysis of Intellectual Disability Copy Number Variants for Association With Schizophrenia. *JAMA psychiatry* **73,** 963–969 (2016).
224. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204,** 108–114 (2014).
225. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. *bioRxiv* 68593 (2016). doi:10.1101/068593
226. Hubbard, L. *et al.* Evidence of common genetic overlap between schizophrenia and cognition. *Schizophr. Bull.* **42,** 832–842 (2016).
227. Pocklington, A. J. *et al.* Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia. *Neuron* **86,** 1203–1214 (2015).
228. Gur, R. E. *et al.* A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. *Mol. Psychiatry* 1–9 (2017). doi:10.1038/mp.2017.161
229. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).
230. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
231. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–60 (2008).
232. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–9 (2006).
233. Loh, P., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48,** 811–6 (2016).
234. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–83 (2016).

235. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
236. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
237. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).
238. Zuvich, R. L. *et al.* Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet. Epidemiol.* **35,** 887–898 (2011).
239. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31,** 1466–8 (2015).
240. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43,** 969–76 (2011).
241. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36,** 7–10 (2010).
242. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis. Media* **35,** (2009).
243. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91,** 1011–21 (2012).
244. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–15 (2013).
245. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
246. Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.* **175,** 315–32 (1948).
247. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 1412–7 (2006).
248. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–22 (2011).
249. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23,** 555–67 (2013).
250. Pradhan, S., Bacolla, A., Wells, R. D. & Roberts, R. J. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J. Biol. Chem.* **274,** 33002–10 (1999).
251. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99,** 247–57 (1999).
252. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129,** 1983–93 (2002).
253. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502,** 472–9 (2013).
254. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–26 (1992).
255. Spiers, H. *et al.* Methylomic trajectories across human fetal brain development. *Genome Res.* **25,** 338–52 (2015).
256. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13,** 1095–107 (1992).
257. Ponger, L., Duret, L. & Mouchiroud, D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* **11,** 1854–60 (2001).
258. Jaffe, A. E. *et al.* Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* **19,** 40–7 (2016).

259. Mill, J. *et al.* Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am. J. Hum. Genet.* **82,** 696–711 (2008).
260. Xiao, Y. *et al.* The DNA methylome and transcriptome of different brain regions in schizophrenia and bipolar disorder. *PLoS One* **9,** e95875 (2014).
261. Wockner, L. F. *et al.* Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Transl. Psychiatry* **4,** e339 (2014).
262. Numata, S., Ye, T., Herman, M. & Lipska, B. K. DNA methylation changes in the postmortem dorsolateral prefrontal cortex of patients with schizophrenia. *Front. Genet.* **5,** 280 (2014).
263. Wockner, L. F. *et al.* Brain-specific epigenetic markers of schizophrenia. *Transl. Psychiatry* **5,** e680 (2015).
264. Viana, J. *et al.* Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions. *Hum. Mol. Genet.* **26,** 210–225 (2017).
265. Lee, S.-A. & Huang, K.-C. Epigenetic profiling of human brain differential DNA methylation networks in schizophrenia. *BMC Med. Genomics* **9,** 68 (2016).
266. Ruzicka, W. B., Subburaju, S. & Benes, F. M. Variability of DNA Methylation within Schizophrenia Risk Loci across Subregions of Human Hippocampus. *Genes (Basel).* **8,** 143 (2017).
267. Dong, E., Gavin, D. P., Chen, Y. & Davis, J. Upregulation of TET1 and downregulation of APOBEC3A and APOBEC3C in the parietal cortex of psychotic patients. *Transl. Psychiatry* **2,** e159 (2012).
268. Zhubi, A. *et al.* An upregulation of DNA-methyltransferase 1 and 3a expressed in telencephalic GABAergic neurons of schizophrenia patients is also detected in peripheral blood lymphocytes. *Schizophr. Res.* **111,** 115–22 (2009).
269. Auta, J. *et al.* DNA-methylation gene network dysregulation in peripheral blood lymphocytes of schizophrenia patients. *Schizophr. Res.* **150,** 312–8 (2013).
270. Nishioka, M. *et al.* Comprehensive DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia. *J. Hum. Genet.* **58,** 91–7 (2013).
271. Ota, V. K. *et al.* Changes in gene expression and methylation in the blood of patients with first-episode psychosis. *Schizophr. Res.* **159,** 358–64 (2014).
272. Kinoshita, M. *et al.* Aberrant DNA methylation of blood in schizophrenia by adjusting for estimated cellular proportions. *Neuromolecular Med.* **16,** 697–703 (2014).
273. Montano, C. *et al.* Association of DNA Methylation Differences With Schizophrenia in an Epigenome-Wide Association Study. *JAMA psychiatry* **73,** 506–14 (2016).
274. Kebir, O. *et al.* Methylomic changes during conversion to psychosis. *Mol. Psychiatry* **22,** 512–518 (2017).
275. Davies, M. N. *et al.* Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* **13,** R43 (2012).
276. Dempster, E. L. *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* **20,** 4786–96 (2011).
277. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10,** 1024–32 (2015).
278. Walton, E. *et al.* Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophr. Bull.* **42,** 406–14 (2016).
279. Atkin, K. *et al.* Neutropenia and agranulocytosis in patients receiving clozapine in the UK and Ireland. *Br. J. Psychiatry* **169,** 483–8 (1996).
280. Reynolds, G. P. & Beasley, C. L. GABAergic neuronal subtypes in the human frontal cortex--development and deficits in schizophrenia. *J. Chem. Neuroanat.* **22,** 95–100 (2001).
281. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association

Studies. *BMC Bioinformatics* **18,** 105 (2017).
282. Kaushal, A. *et al.* Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics* **18,** 216 (2017).
283. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).
284. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15,** R31 (2014).
285. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1,** 239–59 (2009).
286. Bell, J. T. & Spector, T. D. A twin approach to unraveling epigenetics. *Trends Genet.* **27,** 116–25 (2011).
287. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 10604–9 (2005).
288. Castellani, C. A. *et al.* DNA methylation differences in monozygotic twin pairs discordant for schizophrenia identifies psychosis related genes and networks. *BMC Med. Genomics* **8,** 17 (2015).
289. Melka, M. G., Castellani, C. A., O'Reilly, R. & Singh, S. M. Insights into the origin of DNA methylation differences between monozygotic twins discordant for schizophrenia. *J. Mol. psychiatry* **3,** 7 (2015).
290. Lu, J. H., Chung, M. Y., Hwang, B. & Chien, H. P. Monozygotic twins with chromosome 22q11 microdeletion and discordant phenotypes in cardiovascular patterning. *Pediatr. Cardiol.* **22,** 260–3 (2001).
291. Gul, A. *et al.* Prenatal diagnosis of 22q11.2 deletion syndrome in twin pregnancy: a case report. *J. Clin. Ultrasound* **41 Suppl 1,** 6–9 (2013).
292. Melka, M. G., Rajakumar, N., O'Reilly, R. & Singh, S. M. Olanzapine-induced DNA methylation in the hippocampus and cerebellum in genes mapped to human 22q11 and implicated in schizophrenia. *Psychiatr. Genet.* **25,** 88–94 (2015).
293. Loppin, B. *et al.* The histone H3.3 chaperone HIRA is essential for chromatin assembly in the male pronucleus. *Nature* **437,** 1386–90 (2005).
294. Nashun, B. *et al.* Continuous Histone Replacement by Hira Is Essential for Normal Transcriptional Regulation and De Novo DNA Methylation during Mouse Oogenesis. *Mol. Cell* **60,** 611–25 (2015).
295. Han, L., Witmer, P. D., Casey, E., Valle, D. & Sukumar, S. DNA methylation regulates MicroRNA expression. *Cancer Biol. Ther.* **6,** 1284–8 (2007).
296. Sinkkonen, L. *et al.* MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **15,** 259–67 (2008).
297. Benetti, R. *et al.* A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. *Nat. Struct. Mol. Biol.* **15,** 268–79 (2008).
298. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516,** 56–61 (2014).
299. Melton, C., Judson, R. L. & Blelloch, R. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature* **463,** 621–6 (2010).
300. Feng, J. *et al.* Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat. Neurosci.* **13,** 423–30 (2010).
301. Deboer, T., Wu, Z., Lee, A. & Simon, T. J. Hippocampal volume reduction in children with chromosome 22q11.2 deletion syndrome is associated with cognitive impairment. *Behav. Brain Funct.* **3,** 54 (2007).
302. Debbané, M., Schaer, M., Farhoumand, R., Glaser, B. & Eliez, S. Hippocampal volume reduction in 22q11.2 deletion syndrome. *Neuropsychologia* **44,** 2360–5 (2006).
303. Flahault, A., Schaer, M., Ottet, M.-C., Debbané, M. & Eliez, S. Hippocampal volume reduction in chromosome 22q11.2 deletion syndrome (22q11.2DS): a longitudinal study of morphometry and symptomatology. *Psychiatry Res.* **203,** 1–5 (2012).
304. Su, Z., Xia, J. & Zhao, Z. Functional complementation between transcriptional

methylation regulation and post-transcriptional microRNA regulation in the human genome. *BMC Genomics* **12 Suppl 5,** S15 (2011).
305. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41,** e90 (2013).
306. Fortin, J.-P., Triche, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33,** 558–560 (2017).
307. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–9 (2013).
308. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45,** e22 (2017).
309. Zhang, F. F. *et al.* Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics* **6,** 623–9 (2011).
310. Gao, X., Jia, M., Zhang, Y., Breitling, L. P. & Brenner, H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin. Epigenetics* **7,** 113 (2015).
311. Ambatipudi, S. *et al.* Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* **8,** 599–618 (2016).
312. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **120,** 1425–31 (2012).
313. Breton, C. V. *et al.* Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS One* **9,** e99716 (2014).
314. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).
315. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8,** 6 (2015).
316. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
317. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).
318. Riley, J. W. *et al.* The American Soldier: Adjustment During Army Life. *American Sociological Review* **14,** 557 (1949).
319. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32,** 286–8 (2016).
320. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11,** R14 (2010).
321. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **2294,** 00–00 (2015).
322. Breeze, C. E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep.* **17,** 2137–2150 (2016).
323. Germain, R. N. T-cell development and the CD4-CD8 lineage decision. *Nat. Rev. Immunol.* **2,** 309–22 (2002).
324. Bocker, M. T. *et al.* Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood* **117,** e182-9 (2011).
325. Koestler, D. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8,** 816–26 (2013).
326. Shiwa, Y. *et al.* Adjustment of Cell-Type Composition Minimizes Systematic Bias in Blood DNA Methylation Profiles Derived by DNA Collection Protocols. *PLoS One* **11,** e0147519 (2016).

327. Yousefi, P. *et al.* Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environ. Mol. Mutagen.* **56,** 751–8 (2015).
328. Hattab, M. W. *et al.* Correcting for cell-type effects in DNA methylation studies: reference-based method outperforms latent variable approaches in empirical studies. *Genome Biol.* **18,** 24 (2017).
329. Zheng, S. C. *et al.* Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat. Methods* **14,** 216–217 (2017).
330. Strong, E. *et al.* Symmetrical Dose-Dependent DNA-Methylation Profiles in Children with Deletion or Duplication of 7q11.23. *Am. J. Hum. Genet.* **97,** 216–27 (2015).
331. Meyer-Lindenberg, A., Mervis, C. B. & Berman, K. F. Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour. *Nat. Rev. Neurosci.* **7,** 380–93 (2006).
332. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).
333. Middleton, F. a, Mirnics, K., Pierri, J. N., Lewis, D. a & Levitt, P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J. Neurosci.* **22,** 2718–29 (2002).
334. Prabakaran, S. *et al.* Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol. Psychiatry* **9,** 684–97, 643 (2004).
335. Napoli, E. *et al.* Mitochondrial Citrate Transporter-dependent Metabolic Signature in the 22q11.2 Deletion Syndrome. *J. Biol. Chem.* **290,** 23240–53 (2015).
336. Sanders, S. S. *et al.* Curation of the Mammalian Palmitoylome Indicates a Pivotal Role for Palmitoylation in Diseases and Disorders of the Nervous System and Cancers. *PLoS Comput. Biol.* **11,** e1004405 (2015).
337. Amminger, G. P. *et al.* Long-chain omega-3 fatty acids for indicated prevention of psychotic disorders: a randomized, placebo-controlled trial. *Arch. Gen. Psychiatry* **67,** 146–54 (2010).
338. McGorry, P. D. *et al.* Effect of ω-3 Polyunsaturated Fatty Acids in Young People at Ultrahigh Risk for Psychotic Disorders: The NEURAPRO Randomized Clinical Trial. *JAMA psychiatry* **74,** 19–27 (2017).
339. Washbourne, P. *et al.* Cell adhesion molecules in synapse formation. *J. Neurosci.* **24,** 9244–9 (2004).
340. Pocklington, A. J., O'Donovan, M. & Owen, M. J. The synapse in schizophrenia. *Eur. J. Neurosci.* **39,** 1059–67 (2014).
341. Fan, Y. *et al.* Focal adhesion dynamics are altered in schizophrenia. *Biol. Psychiatry* **74,** 418–26 (2013).
342. Maeda, N. Proteoglycans and neuronal migration in the cerebral cortex during development and disease. *Front. Neurosci.* **9,** 98 (2015).
343. Meechan, D. W., Tucker, E. S., Maynard, T. M. & LaMantia, A.-S. Cxcr4 regulation of interneuron migration is disrupted in 22q11.2 deletion syndrome. *Proc. Natl. Acad. Sci.* **109,** 18601–18606 (2012).
344. Schubert, K. O., Föcking, M., Prehn, J. H. M. & Cotter, D. R. Hypothesis review: are clathrin-mediated endocytosis and clathrin-dependent membrane and protein trafficking core pathophysiological processes in schizophrenia and bipolar disorder? *Mol. Psychiatry* **17,** 669–81 (2012).
345. Zhao, Z. *et al.* Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. *Mol. Psychiatry* **20,** 563–572 (2015).
346. Forstner, A. J., Degenhardt, F., Schratt, G. & Nöthen, M. M. MicroRNAs as the cause of schizophrenia in 22q11.2 deletion carriers, and possible implications for idiopathic disease: a mini-review. *Front. Mol. Neurosci.* **6,** 47 (2013).
347. Mendoza, M. C., Er, E. E. & Blenis, J. The Ras-ERK and PI3K-mTOR pathways: cross-talk

and compensation. *Trends Biochem. Sci.* **36,** 320–8 (2011).
348. Reddy, B. V. V. G. & Irvine, K. D. Regulation of Hippo signaling by EGFR-MAPK signaling through Ajuba family proteins. *Dev. Cell* **24,** 459–71 (2013).
349. Emamian, E. S. AKT/GSK3 signaling pathway and schizophrenia. *Front. Mol. Neurosci.* **5,** 33 (2012).
350. Funk, A. J., McCullumsmith, R. E., Haroutunian, V. & Meador-Woodruff, J. H. Abnormal activity of the MAPK- and cAMP-associated signaling pathways in frontal cortical areas in postmortem brain in schizophrenia. *Neuropsychopharmacology* **37,** 896–905 (2012).
351. Moon, A. M. *et al.* Crkl deficiency disrupts Fgf8 signaling in a mouse model of 22q11 deletion syndromes. *Dev. Cell* **10,** 71–80 (2006).
352. Tamura, M., Mukai, J., Gordon, J. A. & Gogos, J. A. Developmental Inhibition of Gsk3 Rescues Behavioral and Neurophysiological Deficits in a Mouse Model of Schizophrenia Predisposition. *Neuron* **89,** 1100–9 (2016).
353. Chelko, S. P. *et al.* Central role for GSK3β in the pathogenesis of arrhythmogenic cardiomyopathy. *JCI insight* **1,** 1–20 (2016).
354. Kerkela, R. *et al.* Deletion of GSK-3beta in mice leads to hypertrophic cardiomyopathy secondary to cardiomyoblast hyperproliferation. *J. Clin. Invest.* **118,** 3609–18 (2008).
355. Koo, B.-K. *et al.* Mind bomb-2 is an E3 ligase for Notch ligand. *J. Biol. Chem.* **280,** 22335–42 (2005).
356. Lasky, J. L. & Wu, H. Notch signaling, brain development, and human disease. *Pediatr. Res.* **57,** 104R–109R (2005).
357. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–84 (2014).
358. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154,** 518–29 (2013).
359. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44,** 1365–9 (2012).
360. Barel, O. *et al.* Deleterious variants in TRAK1 disrupt mitochondrial movement and cause fatal encephalopathy. *Brain* **140,** 568–581 (2017).
361. Ogawa, F. *et al.* DISC1 complexes with TRAK1 and Miro1 to modulate anterograde axonal mitochondrial trafficking. *Hum. Mol. Genet.* **23,** 906–19 (2014).
362. Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9,** 1415–23 (2000).
363. Mao, Y. *et al.* Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of GSK3beta/beta-catenin signaling. *Cell* **136,** 1017–31 (2009).
364. Shu, T., Butz, K. G., Plachez, C., Gronostajski, R. M. & Richards, L. J. Abnormal development of forebrain midline glia and commissural projections in Nfia knock-out mice. *J. Neurosci.* **23,** 203–12 (2003).
365. Campbell, C. E. *et al.* The transcription factor Nfix is essential for normal brain development. *BMC Dev. Biol.* **8,** 52 (2008).
366. Lu, W. *et al.* NFIA haploinsufficiency is associated with a CNS malformation syndrome and urinary tract defects. *PLoS Genet.* **3,** e80 (2007).
367. Dolan, M. *et al.* A novel microdeletion/microduplication syndrome of 19p13.13. *Genet. Med.* **12,** 503–11 (2010).
368. Lesch, K.-P. *et al.* Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Mol. Psychiatry* **16,** 491–503 (2011).
369. Carter, M. G. *et al.* Mice deficient in the candidate tumor suppressor gene Hic1 exhibit developmental defects of structures affected in the Miller-Dieker syndrome. *Hum. Mol. Genet.* **9,** 413–9 (2000).
370. Smith, T. M., Wang, X., Zhang, W., Kulyk, W. & Nazarali, A. J. Hoxa2 plays a direct role in murine palate development. *Dev. Dyn.* **238,** 2364–73 (2009).
371. Faqeih, E. A. *et al.* Novel homozygous DEAF1 variant suspected in causing white matter disease, intellectual disability, and microcephaly. *Am. J. Med. Genet. A* **164A,** 1565–70

(2014).
372. Vulto-van Silfhout, A. T. *et al.* Mutations affecting the SAND domain of DEAF1 cause intellectual disability with severe speech impairment and behavioral problems. *Am. J. Hum. Genet.* **94,** 649–61 (2014).
373. Rossi, M. R. *et al.* Clinical and genomic characterization of distal duplications and deletions of chromosome 4q: study of two cases and review of the literature. *Am. J. Med. Genet. A* **149A,** 2788–94 (2009).
374. Zhang, Q. *et al.* Impaired Dendritic Development and Memory in Sorbs2 Knock-Out Mice. *J. Neurosci.* **36,** 2247–60 (2016).
375. Park, H. *et al.* Mice lacking the PSD-95-interacting E3 ligase, Dorfin/Rnf19a, display reduced adult neurogenesis, enhanced long-term potentiation, and impaired contextual fear conditioning. *Sci. Rep.* **5,** 16410 (2015).
376. Medrihan, L. *et al.* Neurobeachin, a protein implicated in membrane protein traffic and autism, is required for the formation and functioning of central synapses. *J. Physiol.* **587,** 5095–106 (2009).
377. Castermans, D. *et al.* The neurobeachin gene is disrupted by a translocation in a patient with idiopathic autism. *J. Med. Genet.* **40,** 352–6 (2003).
378. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.* **2,** 21–32 (2001).
379. Maynard, T. M. *et al.* No evidence for parental imprinting of mouse 22q11 gene orthologs. *Mamm. Genome* **17,** 822–32 (2006).
380. European Network of National Networks studying Gene-Environment Interactions in Schizophrenia (EU-GEI) *et al.* Identifying gene-environment interactions in schizophrenia: contemporary challenges for integrated, large-scale investigations. *Schizophr. Bull.* **40,** 729–36 (2014).
381. Shiohama, A., Sasaki, T., Noda, S., Minoshima, S. & Shimizu, N. Molecular cloning and expression analysis of a novel gene DGCR8 located in the DiGeorge syndrome chromosomal region. *Biochem. Biophys. Res. Commun.* **304,** 184–90 (2003).
382. Nilsen, T. W. Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends Genet.* **23,** 243–9 (2007).
383. Miyoshi, K., Miyoshi, T. & Siomi, H. Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. *Mol. Genet. Genomics* **284,** 95–103 (2010).
384. Shiohama, A., Sasaki, T., Noda, S., Minoshima, S. & Shimizu, N. Nucleolar localization of DGCR8 and identification of eleven DGCR8-associated proteins. *Exp. Cell Res.* **313,** 4196–207 (2007).
385. Dai, L. *et al.* Cytoplasmic Drosha activity generated by alternative splicing. *Nucleic Acids Res.* **44,** 10454–10466 (2016).
386. Yeom, K.-H., Lee, Y., Han, J., Suh, M. R. & Kim, V. N. Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. *Nucleic Acids Res.* **34,** 4622–9 (2006).
387. Barr, I. *et al.* DiGeorge critical region 8 (DGCR8) is a double-cysteine-ligated heme protein. *J. Biol. Chem.* **286,** 16716–25 (2011).
388. Quick-Cleveland, J. *et al.* The DGCR8 RNA-binding heme domain recognizes primary microRNAs by clamping the hairpin. *Cell Rep.* **7,** 1994–2005 (2014).
389. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15,** 509–24 (2014).
390. Macias, S. *et al.* DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nat. Struct. Mol. Biol.* **19,** 760–6 (2012).
391. Kadener, S. *et al.* Genome-wide identification of targets of the drosha-pasha/DGCR8 complex. *RNA* **15,** 537–545 (2009).
392. Seong, Y. *et al.* Global identification of target recognition and cleavage by the Microprocessor in human ES cells. *Nucleic Acids Res.* **42,** 12806–21 (2014).
393. Johanson, T. M. *et al.* Drosha controls dendritic cell development by cleaving

messenger RNAs encoding inhibitors of myelopoiesis. *Nat. Immunol.* **16,** 1134–41 (2015).
394. Bartram, M. P. *et al.* Loss of Dgcr8-mediated microRNA expression in the kidney results in hydronephrosis and renal malformation. *BMC Nephrol.* **16,** 55 (2015).
395. Rao, P. K. *et al.* Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circ. Res.* **105,** 585–94 (2009).
396. Knuckles, P. *et al.* Drosha regulates neurogenesis by controlling neurogenin 2 expression independent of microRNAs. *Nat. Neurosci.* **15,** 962–9 (2012).
397. Marinaro, F. *et al.* MicroRNA-independent functions of DGCR8 are essential for neocortical development and TBR1 expression. *EMBO Rep.* **18,** 603–618 (2017).
398. Ouchi, Y. *et al.* Reduced adult hippocampal neurogenesis and working memory deficits in the Dgcr8-deficient mouse model of 22q11.2 deletion-associated schizophrenia can be rescued by IGF2. *J. Neurosci.* **33,** 9408–19 (2013).
399. Hsu, R. *et al.* Loss of microRNAs in pyramidal neurons leads to specific changes in inhibitory synaptic transmission in the prefrontal cortex. *Mol. Cell. Neurosci.* **50,** 283–92 (2012).
400. Fénelon, K. *et al.* Deficiency of Dgcr8, a gene disrupted by the 22q11.2 microdeletion, results in altered short-term plasticity in the prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 4447–52 (2011).
401. Toritsuka, M. *et al.* Deficits in microRNA-mediated Cxcr4/Cxcl12 signaling in neurodevelopmental deficits in a 22q11 deletion syndrome mouse model. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17552–7 (2013).
402. Chun, S. *et al.* Specific disruption of thalamic inputs to the auditory cortex in schizophrenia models. *Science* **344,** 1178–82 (2014).
403. Horie, K. *et al.* A homozygous mutant embryonic stem cell bank applicable for phenotype-driven genetic screening. *Nat. Methods* **8,** 1071–7 (2011).
404. Wang, Y. *et al.* Embryonic stem cell-specific microRNAs regulate the G1-S transition and promote rapid proliferation. *Nat. Genet.* **40,** 1478–83 (2008).
405. Ma, Y. *et al.* Functional screen reveals essential roles of miR-27a/24 in differentiation of embryonic stem cells. *EMBO J.* **34,** 361–78 (2015).
406. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).
407. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–23 (2013).
408. Arber, C. *et al.* Activin A directs striatal projection neuron differentiation of human pluripotent stem cells. *Development* **142,** 1375–86 (2015).
409. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One* **10,** 1–11 (2015).
410. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* 1–8 (2015). doi:10.1038/nmeth.3284
411. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42,** e168–e168 (2014).
412. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–8 (2001).
413. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).
414. Triboulet, R., Chang, H.-M., Lapierre, R. J. & Gregory, R. I. Post-transcriptional control of DGCR8 expression by the Microprocessor. *RNA* **15,** 1005–11 (2009).
415. Han, J. *et al.* Posttranscriptional crossregulation between Drosha and DGCR8. *Cell* **136,** 75–84 (2009).
416. Schofield, C. M. *et al.* Monoallelic deletion of the microRNA biogenesis gene Dgcr8 produces deficits in the development of excitatory synaptic transmission in the

prefrontal cortex. *Neural Dev.* **6,** 11 (2011).
417. Martin, R. *et al.* A Drosophila pasha mutant distinguishes the canonical microRNA and mirtron pathways. *Mol Cell Biol* **29,** 861–870 (2009).
418. Stark, K. L. *et al.* Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat. Genet.* **40,** 751–60 (2008).
419. Zhao, D. *et al.* MicroRNA Profiling of Neurons Generated Using Induced Pluripotent Stem Cells Derived from Patients with Schizophrenia and Schizoaffective Disorder, and 22q11.2 Del. *PLoS One* **10,** e0132387 (2015).
420. Chen, Z. *et al.* DiGeorge syndrome critical region 8 (DGCR8) protein-mediated microRNA biogenesis is essential for vascular smooth muscle cell development in mice. *J. Biol. Chem.* **287,** 19018–28 (2012).
421. Mao, S. *et al.* miR-17-92 facilitates neuronal differentiation of transplanted neural stem/precursor cells under neuroinflammatory conditions. *J. Neuroinflammation* **13,** 208 (2016).
422. Bian, S. *et al.* MicroRNA cluster miR-17-92 regulates neural stem cell expansion and transition to intermediate progenitors in the developing mouse neocortex. *Cell Rep.* **3,** 1398–1406 (2013).
423. Jin, J. *et al.* miR-17-92 Cluster Regulates Adult Hippocampal Neurogenesis, Anxiety, and Depression. *Cell Rep.* **16,** 1653–63 (2016).
424. Knuckles, P. *et al.* Drosha regulates neurogenesis by controlling Neurogenin 2 expression independent of microRNAs. *Nat. Neurosci.* **15,** 962–969 (2012).
425. Di Carlo, V. *et al.* TDP-43 regulates the microprocessor complex activity during in vitro neuronal differentiation. *Mol. Neurobiol.* **48,** 952–963 (2013).
426. Zhang, Y. *et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78,** 785–98 (2013).
427. Thoma, E. C. *et al.* Ectopic expression of neurogenin 2 alone is sufficient to induce differentiation of embryonic stem cells into mature neurons. *PLoS One* **7,** (2012).
428. Ho, S. M. *et al.* Rapid Ngn2-induction of excitatory neurons from hiPSC-derived neural progenitor cells. *Methods* **101,** 113–124 (2016).
429. Sansom, S. N. *et al.* The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis. *PLoS Genet.* **5,** e1000511 (2009).
430. Kadener, S. *et al.* Genome-wide identification of targets of the drosha-pasha/DGCR8 complex. *RNA* **15,** 537–45 (2009).
431. Shenoy, A. & Blelloch, R. Genomic analysis suggests that mRNA destabilization by the microprocessor is specialized for the auto-regulation of Dgcr8. *PLoS One* **4,** e6971 (2009).
432. Seong, Y. *et al.* Global identification of target recognition and cleavage by the Microprocessor in human ES cells. *Nucleic Acids Res.* **42,** 12806–21 (2014).
433. Kim, B., Jeong, K. & Kim, V. N. Genome-wide Mapping of DROSHA Cleavage Sites on Primary MicroRNAs and Noncanonical Substrates. *Mol. Cell* **66,** 258–269.e5 (2017).
434. Kiss, T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.* **20,** 3617–22 (2001).
435. Macias, S., Cordiner, R. A., Gautier, P., Plass, M. & Cáceres, J. F. DGCR8 Acts as an Adaptor for the Exosome Complex to Degrade Double-Stranded Structured RNAs. *Mol. Cell* **60,** 873–85 (2015).
436. Gromak, N. *et al.* Drosha regulates gene expression independently of RNA cleavage function. *Cell Rep.* **5,** 1499–510 (2013).
437. Havens, M. A., Reich, A. A. & Hastings, M. L. Drosha promotes splicing of a pre-microRNA-like alternative exon. *PLoS Genet.* **10,** e1004312 (2014).
438. Melamed, Z. *et al.* Alternative splicing regulates biogenesis of miRNAs located across exon-intron junctions. *Mol. Cell* **50,** 869–81 (2013).
439. Agranat-Tamir, L., Shomron, N., Sperling, J. & Sperling, R. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. *Nucleic Acids Res.* **42,**

4640–51 (2014).
440. van Beveren, N. J. M. *et al.* Functional gene-expression analysis shows involvement of schizophrenia-relevant pathways in patients with 22q11 deletion syndrome. *PLoS One* **7,** e33473 (2012).
441. Jalbrzikowski, M. *et al.* Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder. *PLoS One* **10,** e0132542 (2015).
442. Shi, Y., Inoue, H., Wu, J. C. & Yamanaka, S. Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.* **16,** 115–130 (2017).
443. Tai, D. J. C. *et al.* Engineering microdeletions and microduplications by targeting segmental duplications with CRISPR. *Nat. Neurosci.* **19,** 517–22 (2016).
444. Pedrosa, E. *et al.* Development of patient-specific neurons in schizophrenia using induced pluripotent stem cells. *J. Neurogenet.* **25,** 88–103 (2011).
445. Lin, M. *et al.* Integrative transcriptome network analysis of iPSC-derived neurons from schizophrenia and schizoaffective disorder patients with 22q11.2 deletion. *BMC Syst. Biol.* **10,** 105 (2016).
446. Toyoshima, M. *et al.* Analysis of induced pluripotent stem cells carrying 22q11.2 deletion. *Transl. Psychiatry* **6,** e934 (2016).
447. Bundo, M. *et al.* Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* **81,** 306–13 (2014).
448. Heras, S. R. *et al.* The Microprocessor controls the activity of mammalian retrotransposons. *Nat. Struct. Mol. Biol.* **20,** 1173–81 (2013).
449. Marinaro, F. *et al.* MicroRNA-independent functions of DGCR8 are essential for neocortical development and TBR1 expression. *EMBO Rep.* **18,** 603–618 (2017).
450. Davis, M. P. a *et al.* Large-scale identification of microRNA targets in murine Dgcr8-deficient embryonic stem cell lines. *PLoS One* **7,** e41762 (2012).
451. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–23 (2008).
452. Gu, K.-L. *et al.* Pluripotency-associated miR-290/302 family of microRNAs promote the dismantling of naive pluripotency. *Cell Res.* **26,** 350–66 (2016).
453. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–20 (2014).
454. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinforma.* http://www.bioinformatics.babraham.ac.uk/projects/ (2010). doi:citeulike-article-id:11583827
455. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
456. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
457. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
458. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16,** 284–7 (2012).
459. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–6 (2013).
460. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–32 (2013).
461. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24,** 132–41 (2014).
462. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12,** 237–43, 1 p following 243 (2015).
463. Pietersen, C. Y. *et al.* Molecular profiles of pyramidal neurons in the superior temporal cortex in schizophrenia. *J. Neurogenet.* **28,** 53–69 (2014).
464. Maitra, A. *et al.* Genomic alterations in cultured human embryonic stem cells. *Nat.*

Genet. **37,** 1099–103 (2005).
465. Närvä, E. *et al.* High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat. Biotechnol.* **28,** 371–7 (2010).
466. Laurent, L. C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8,** 106–18 (2011).
467. Hussein, S. M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471,** 58–62 (2011).
468. Lefort, N. *et al.* Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. *Nat. Biotechnol.* **26,** 1364–6 (2008).
469. International Stem Cell Initiative *et al.* Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nat. Biotechnol.* **29,** 1132–44 (2011).
470. Nguyen, H. T. *et al.* Gain of 20q11.21 in human embryonic stem cells improves cell survival by increased expression of Bcl-xL. *Mol. Hum. Reprod.* **20,** 168–77 (2014).
471. Garitaonandia, I. *et al.* Increased risk of genetic and epigenetic instability in human embryonic stem cells associated with specific culture conditions. *PLoS One* **10,** e0118307 (2015).
472. Bieging, K. T., Mello, S. S. & Attardi, L. D. Unravelling mechanisms of p53-mediated tumour suppression. *Nat. Rev. Cancer* **14,** 359–70 (2014).
473. Amir, H. *et al.* Spontaneous Single-Copy Loss of TP53 in Human Embryonic Stem Cells Markedly Increases Cell Proliferation and Survival. *Stem Cells* **35,** 872–885 (2017).
474. Francia, S. *et al.* Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature* **488,** 231–5 (2012).
475. Francia, S., Cabrini, M., Matti, V., Oldani, A. & d'Adda di Fagagna, F. DICER, DROSHA and DNA damage response RNAs are necessary for the secondary recruitment of DNA damage response factors. *J. Cell Sci.* **129,** 1468–76 (2016).
476. Wan, G., Mathur, R., Hu, X., Zhang, X. & Lu, X. miRNA response to DNA damage. *Trends Biochem. Sci.* **36,** 478–84 (2011).
477. Suzuki, H. I. *et al.* Modulation of microRNA processing by p53. *Nature* **460,** 529–33 (2009).
478. Dallosso, A. R. *et al.* Frequent long-range epigenetic silencing of protocadherin gene clusters on chromosome 5q31 in Wilms' tumor. *PLoS Genet.* **5,** e1000745 (2009).
479. Coolen, M. W. *et al.* Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. *Nat. Cell Biol.* **12,** 235–46 (2010).
480. Brzustowicz, L. M. & Bassett, A. S. miRNA-mediated risk for schizophrenia in 22q11.2 deletion syndrome. *Front. Genet.* **3,** 291 (2012).
481. Triboulet, R., Chang, H.-M., Lapierre, R. J. & Gregory, R. I. Post-transcriptional control of DGCR8 expression by the Microprocessor. *RNA* **15,** 1005–11 (2009).
482. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel).* **5,** 3 (2016).
483. Starnawska, A. *et al.* Differential DNA methylation at birth associated with mental disorder in individuals with 22q11.2 deletion syndrome. *Transl. Psychiatry* **7,** e1221 (2017).
484. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12,** R10 (2011).
485. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19,** 48–54 (2016).
486. Hannon, E. *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* **17,** 176 (2016).
487. Merico, D. *et al.* MicroRNA Dysregulation, Gene Networks, and Risk for Schizophrenia in

22q11.2 Deletion Syndrome. *Front. Neurol.* **5,** 1–12 (2014).
488. Chakraborty, S., Mehtab, S., Patwardhan, A. & Krishnan, Y. Pri-miR-17-92a transcript folds into a tertiary structure and autoregulates its processing. *Rna* **18,** 1014–1028 (2012).
489. De Decker, H. P. & Lawrenson, J. B. The 22q11.2 deletion: from diversity to a single gene theory. *Genet. Med.* **3,** 2–5 (2001).
490. Chen, L., Zhou, W., Zhang, L. & Zhang, F. Genome architecture and its roles in human copy number variation. *Genomics Inform.* **12,** 136–44 (2014).
491. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
492. McDonald-McGinn, D. M. *et al.* 22q11.2 deletion syndrome. *Nat. Rev. Dis. Prim.* **1,** 15071 (2015).
493. Arguello, P. A., Markx, S., Gogos, J. a & Karayiorgou, M. Development of animal models for schizophrenia. *Dis. Model. Mech.* **3,** 22–6 (2013).
494. Karayiorgou, M., Simon, T. J. & Gogos, J. a. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. *Nat. Rev. Neurosci.* **11,** 402–16 (2010).
495. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–7 (2010).
496. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38,** 23–38 (2013).
497. Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–6 (2015).
498. Jaffe, A. E. & Kleinman, J. E. Genetic and epigenetic analysis of schizophrenia in blood—a no-brainer? *Genome Med.* **8,** 96 (2016).
499. Aberg, K. A. *et al.* Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. *JAMA psychiatry* **71,** 255–64 (2014).
500. Germain, N., Banda, E. & Grabel, L. Embryonic stem cell neurogenesis and neural specification. *J. Cell. Biochem.* **111,** 535–42 (2010).
501. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5,** 16923 (2015).

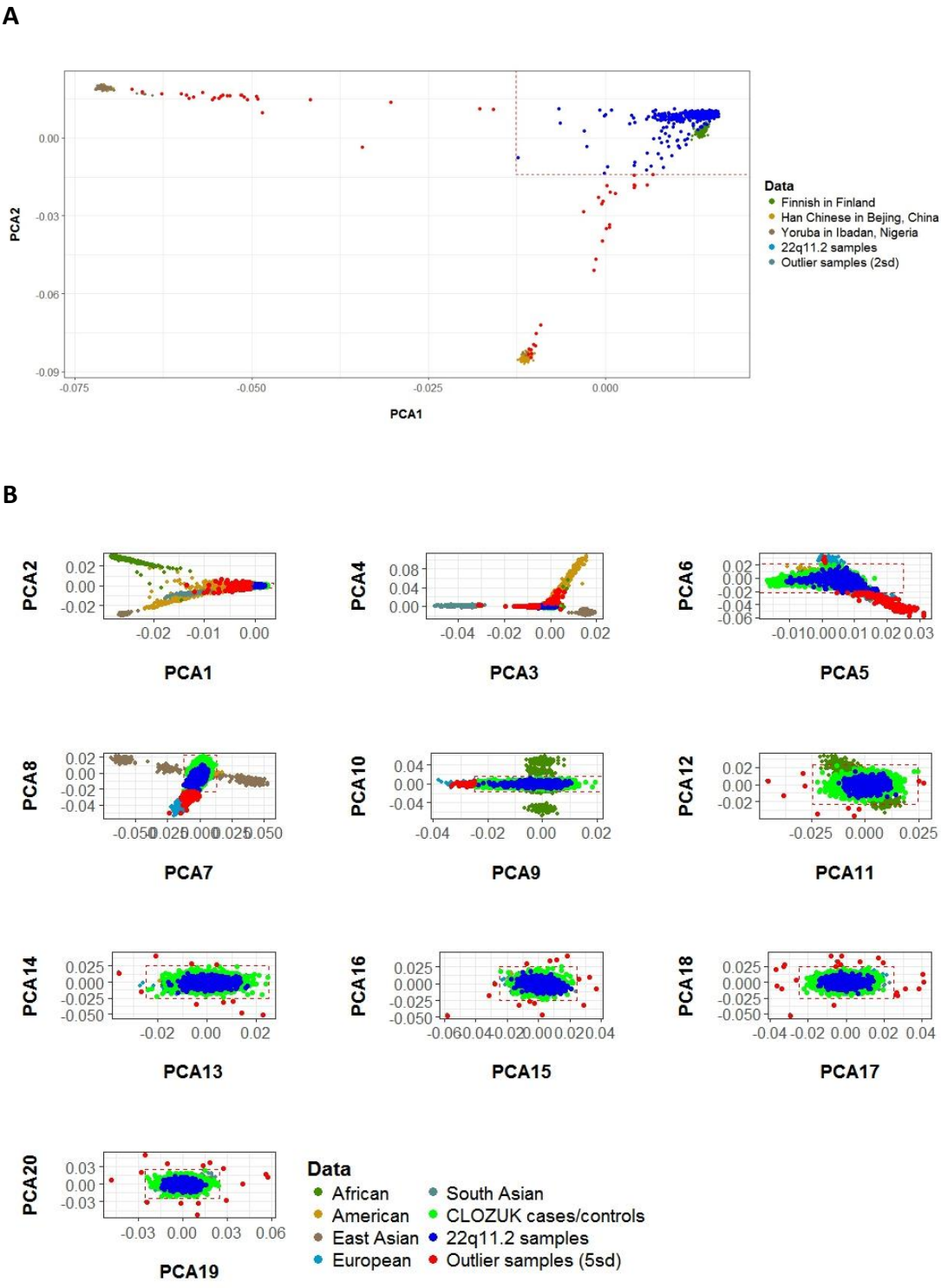
Supplementary information: Schizophrenia polygenic risk score analysis in 22q11.2 deletion syndrome patients

	22q11 SCZ	22q NonAffect	Total
Toronto, Canada	33	45	78
Maastricht, Netherlands	28	29	57
Philadelphia, PA, USA	11	30	41
Leuven, Belgium	8	25	33
Cardiff, UK / London, UK / Dublin, Ireland	10	22	32
Los Angeles, CA, USA	1	11	12
Geneva, Switzerland	7	1	8
Newcastle, Australia	2	5	7
Marseilles, France	0	6	6
Bronx, NY, USA	3	3	6
Durham, NC, USA	1	3	4
Utrecht, Netherlands	2	0	2
Mallorca, Spain	1	0	1
Atlanta, GA, USA	0	1	1
Total	107	181	288

Supplementary table 2-1. Case/Control sample sizes (after quality control) and their origin (international sites participating in the 22q11.2 IBBC).

Step	22q11 samples removed	CLOZUK samples removed	SNP removed	# 22q11.2 samples after step	# CLOZUK samples after step	# SNPs after step
Sex check	**25**	x	x	**1135**	x	**908440**
Exclude non-autosomal SNPs	x	x	**37162**	**1103**	x	**871278**
Exclude SNPs with no name ("00") or with a null allele (0) and for duplicate SNP, removed the one with highest missingness	x	x	**4552**	**1103**	x	**866726**
Exclude SNP with missingness > 5%	x	x	**73976**	**1103**	x	**792750**
Exclude samples with missingness > 3%	**32**	x	x	**1071**	x	**783213**
Perform IBD analysis; for each IBD pair, remove one sample based on phenotype status and sample missingness (PI_HAT > 0.125)	**145**	x	x	**926**	x	**783213**
Merge samples with 1000 genome data, perform PCA and exclude outliers more than 3 standard deviation away from any principal component mean	**56**	x	x	**926**	x	**783213**
Exclude sample with heterozygosity threshold |F| > 0.2	**0**	x	x	**870**	x	**783213**
Exclude SNP with Hardy-Weinberg equilibrium HWE < 10e-5	x	x	**9537**	**870**	x	**783213**
Perform imputation on the Michigan Imputation server	x	x	x	**870**	x	
Perform pre-QC on 22Q imputed data (INFO SCORE 0.9, SNP missingness < 1%, HWE midp < 10e-5)	x	x	**x**	**870**	x	**6346287**
Perform pre-QC on CLOZUK imputed data (INFO SCORE 0.9, SNP missingness 0%, HWE midp < 10e-5)	x	x	x	**870**	**35802**	**199619**
Merge 22Q data with CLOZUK data (keep only common SNPs)	x	x	x	**870**	**35802**	**181030**
Exclude SNP with missingness > 1%	x	x	**2**	**870**	**35802**	**181028**
Exclude samples with missingness > 1% (safety check)	0	0	x	**870**	**35802**	**181028**
Perform IBD analysis; for each IBD pair, remove one sample based on phenotype status and sample missingness (PI_HAT > 0.125)	0	**136**	x	**870**	**35666**	**181028**
Merge samples with 1000 genome data, perform PCA and exclude outliers more than 6 standard deviation away from any principal component mean	**201**	**290**	x	**669**	**35376**	**181028**
Remove samples from CLOZUK that have a known CNV	x	**324**	x	**669**	**35052**	**181028**
Remove variants within 22q11.2 deletion	x	x	**112**	**669**	**35052**	**180916**
Remove variants with MAF < 0.1	x	x	**27021**	**669**	**35052**	**153895**

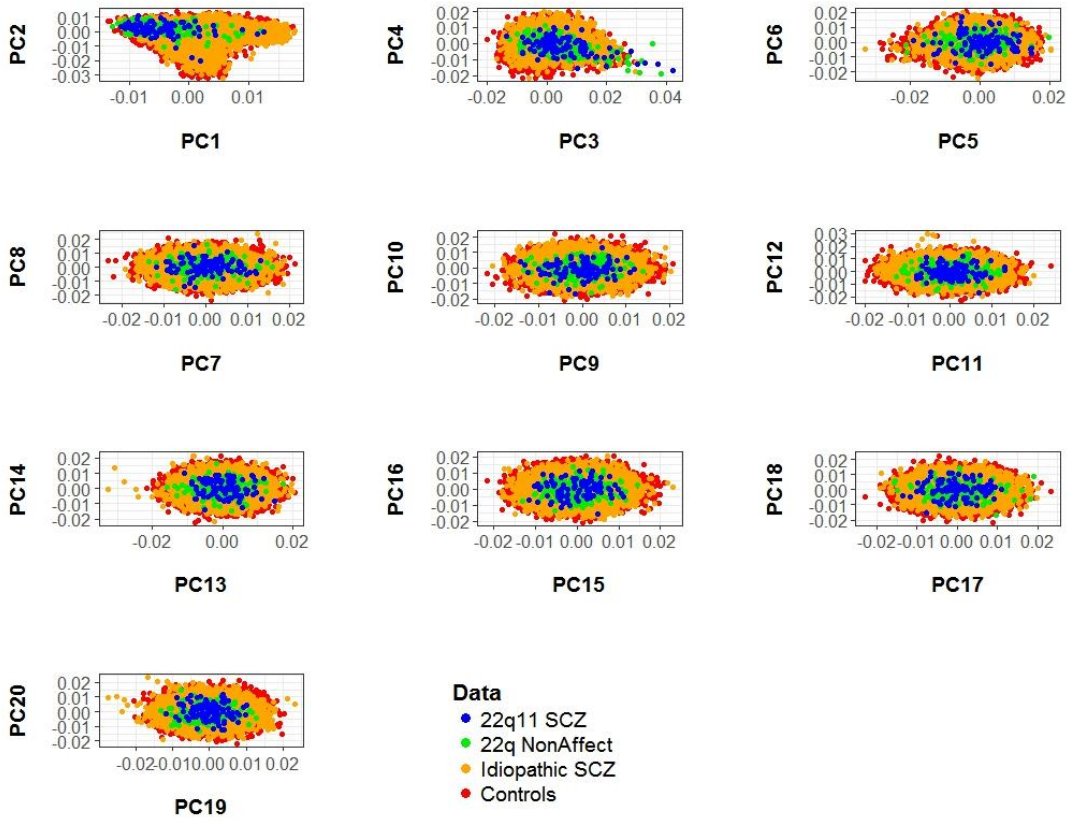
Supplementary table 2-2. Data processing and quality control steps. The different steps that have been performed to process the data before analysis are detailed, as well as the number of samples and variants (single nucleotide polymorphisms, SNPs) removed and remaining after each step.



Supplementary figure 2-1. Principal component analysis (PCA) and ethnicity-based outlier removal before (A) and after (B) imputation and merging 22q11.2 data with CLOZUK/WTCCC data. In each case outliers were defined as samples further away than N standard deviation from the mean of any principal component (N = 2 for (A) and N = 5 for (B)). Red dotted lines delimit this zone of exclusion. PCA was performed after merging data with 1000 Genomes data that contains samples from ethnically different populations and sub-populations. Details of these populations are available in supplementary figure 3).

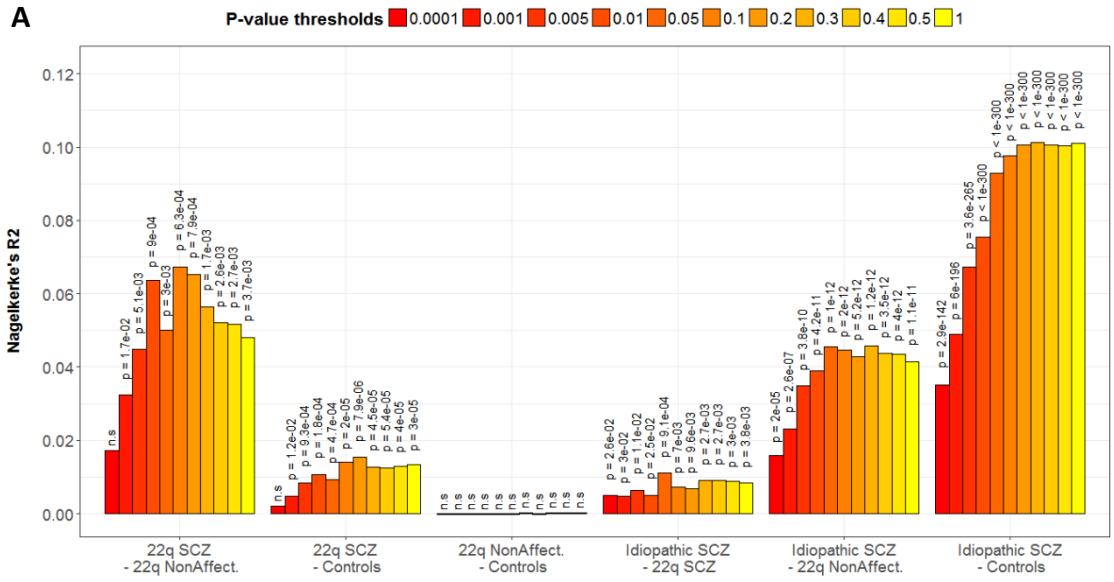
ID	Population	Subpopulation	Size
afr_acb	African	African Caribbeans in Barbados	96
afr_asw	African	Americans of African Ancestry in SW USA	61
afr_esn	African	Esan in Nigeria	99
afr_gwd	African	Gambian in Western Divisions in the Gambia	113
afr_lwk	African	Luhya in Webuye, Kenya	99
afr_msl	African	Mende in Sierra Leone	85
afr_yri	African	Yoruba in Ibadan, Nigeria	108
amr_clm	American	Colombians from Medellin, Colombia	94
amr_mxl	American	Mexican Ancestry from Los Angeles USA	64
amr_pel	American	Peruvians from Lima, Peru	85
amr_pur	American	Puerto Ricans from Puerto Rico	104
eas_cdx	East Asian	Chinese Dai in Xishuangbanna, China	93
eas_chb	East Asian	Han Chinese in Bejing, China	103
eas_chs	East Asian	Southern Han Chinese	105
eas_jpt	East Asian	Japanese in Tokyo, Japan	104
eas_khv	East Asian	Kinh in Ho Chi Minh City, Vietnam	99
eur_ceu	European	Utah Residents (CEPH) with Northern and Western Ancestry	99
eur_fin	European	Finnish in Finland	99
eur_gbr	European	British in England and Scotland	91
eur_ibs	European	Iberian Population in Spain	107
eur_tsi	European	Toscani in Italia	107
sas_beb	South Asian	Bengali from Bangladesh	86
sas_gih	South Asian	Gujarati Indian from Houston, Texas	103
sas_itu	South Asian	Indian Telugu from the UK	102
sas_pjl	South Asian	Punjabi from Lahore, Pakistan	96
sas_stu	South Asian	Sri Lankan Tamil from the UK	102

Supplementary table 2-3. Population and subpopulation details of the 1000 Genomes Project data used in the principal component analysis. Data from the International Genome Sample Resource (IGSR, REF). ID: 1000 Genomes Project Phase 3 population ID; Size: number of samples included (total: 2504 samples).



Supplementary figure 2-2. Principal component analysis (PCA) of the data used for analysis for all principal components used as covariates in the regression analysis model. PCA is showing that the data is relatively homogenous.

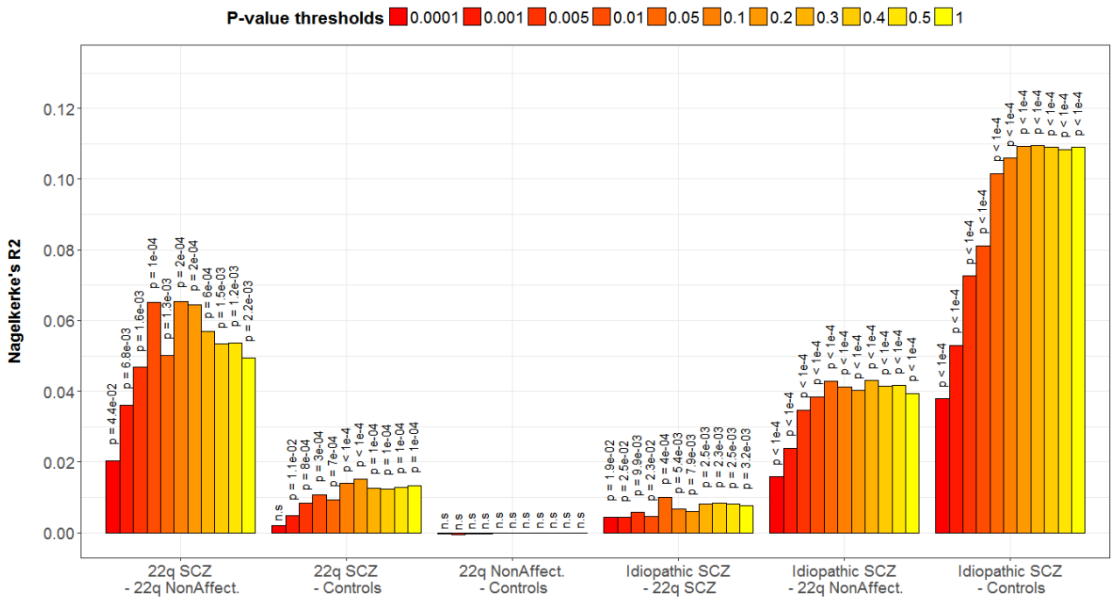
A



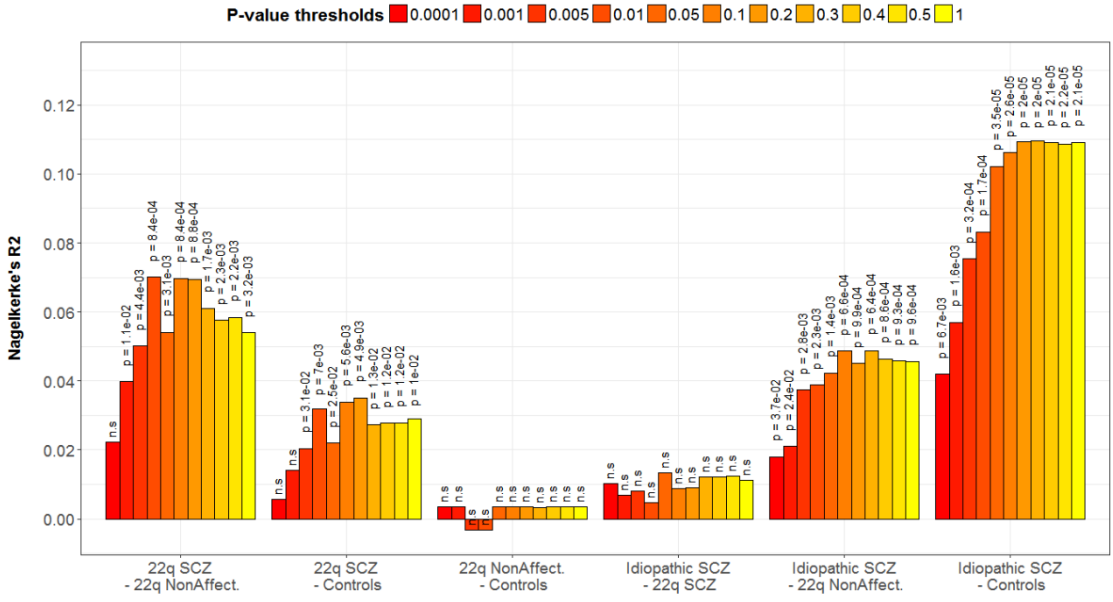
B

	22q11 SCZ	22q NonAffect	Idiopathic SCZ	Controls
Sample size	105	128	5909	22716
Gender ratio M/F	0.91	0.91	0.91	0.91

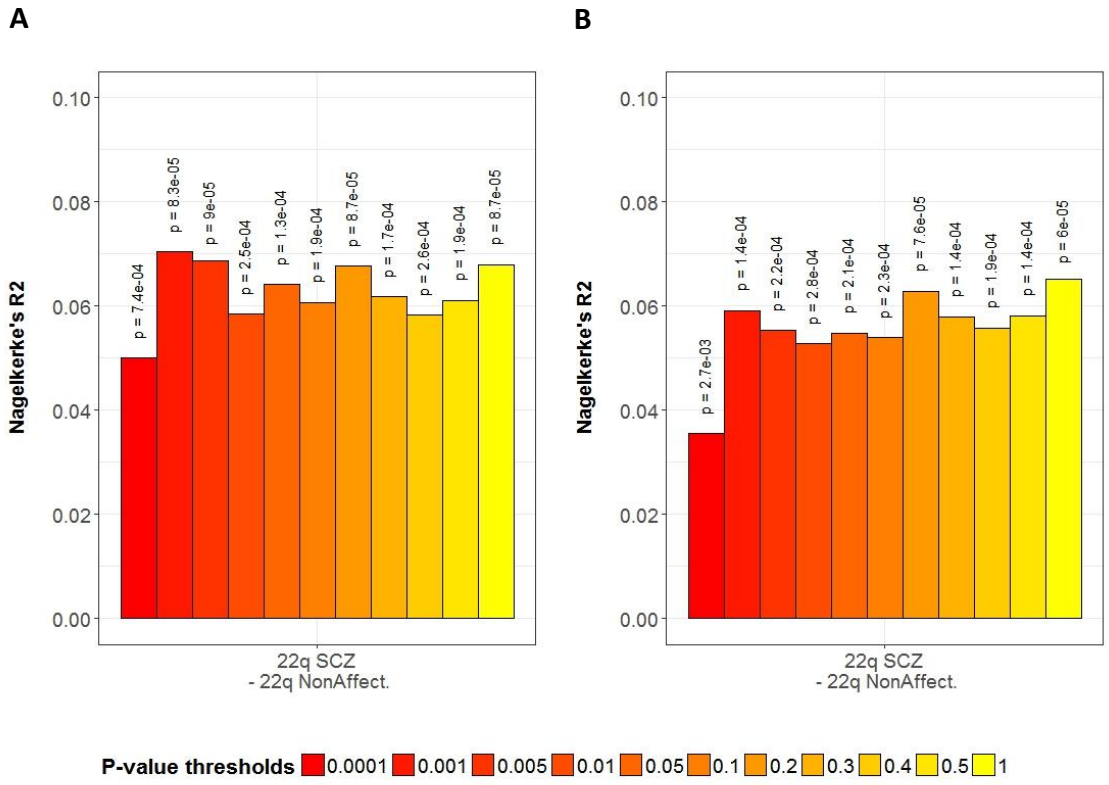
Supplementary figure 2-3. Effect of adjusting the gender ratios. (A) Proportion of variance in schizophrenia explained by polygenic risk score (PRS) between groups with similar gender ratios. Samples were randomly excluded to adjust the gender ratio between groups (N = 10,000 permutations). Gender ratios are adjusted to the ratio of the smallest group (22q11 SCZ) to maximize sample size. Separate analysis for genders was not performed due to small sample sizes. The displayed R^2 values are the mean values of R^2 values for 10,000 permutations. The p-value thresholds pT for selecting risk alleles are shown by colours (cf. legend on top). The displayed p-values on top of bars are the geometric mean values of p-values of association of phenotype with PRS for the 10,000 permutations. **(B) Groups sample sizes after gender ratio adjustment.**



Supplementary figure 2-4. Association of polygenic risk score with phenotype determined by permutation analysis (N = 10,000 permutations). The p-value thresholds pT for selecting risk alleles are shown by colours (cf. legend on top). For each pairwise comparison, the phenotypes were randomised while keeping the same case/control ratio. The p-values represent the probability of obtaining a smaller p-value by chance (after random sampling). Sample sizes and R^2 values obtained are the same than used in Figure 2-4. $p < 1e^{-4}$ is the minimal displayed p-value due to the number of permutation performed (10,000).



Supplementary figure 2-5. Proportion of variance in schizophrenia explained by polygenic risk score (PRS) between groups of equal sample sizes (N = 107 samples, random sample selection, 10,000 permutations). The displayed R^2 values are the mean values of R^2 values for 10,000 permutations. The p-value thresholds pT for selecting risk alleles are shown by colours (cf. legend on top). The displayed p-values on top of bars are the geometric mean values of p-values of association of phenotype with PRS for the 10,000 permutations.

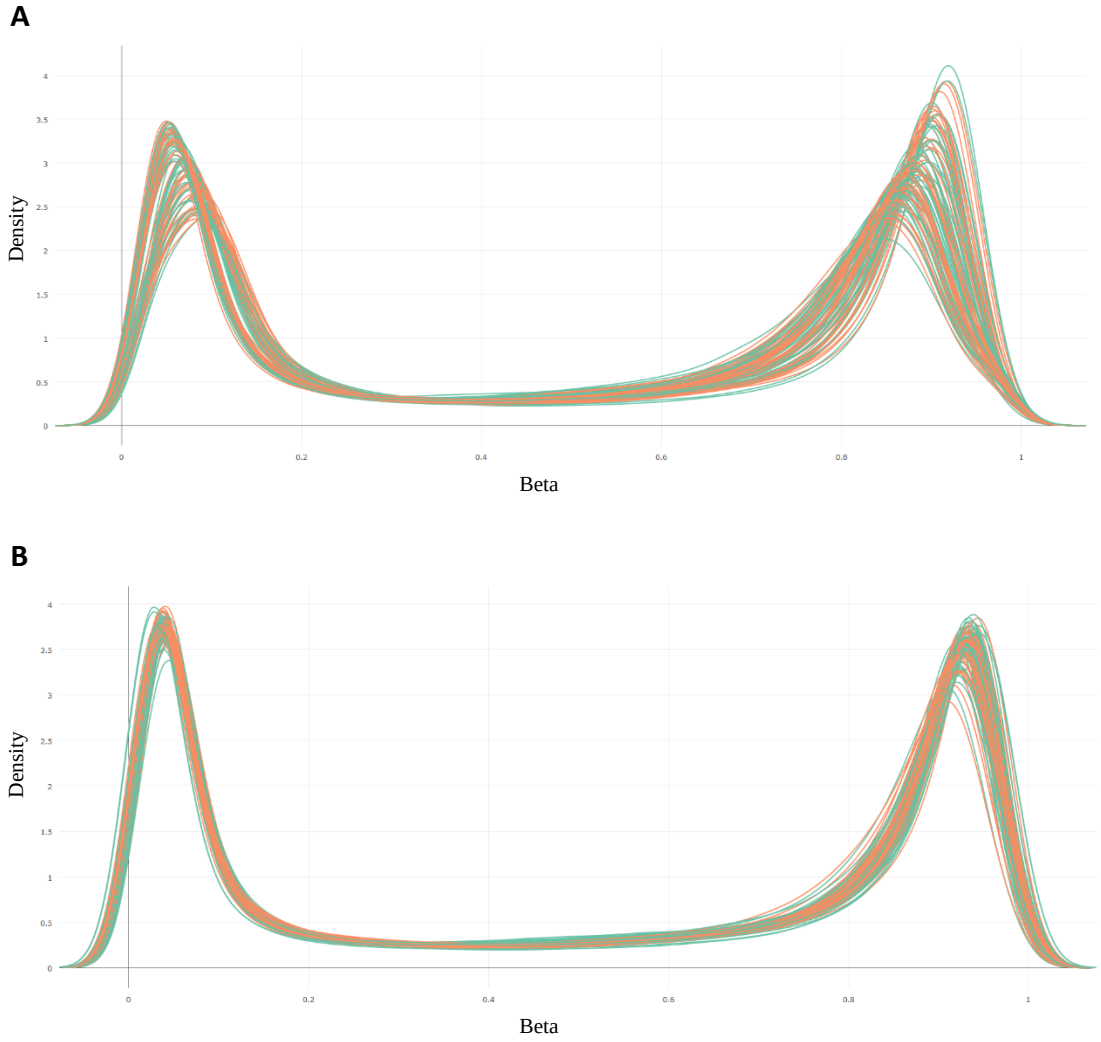


C

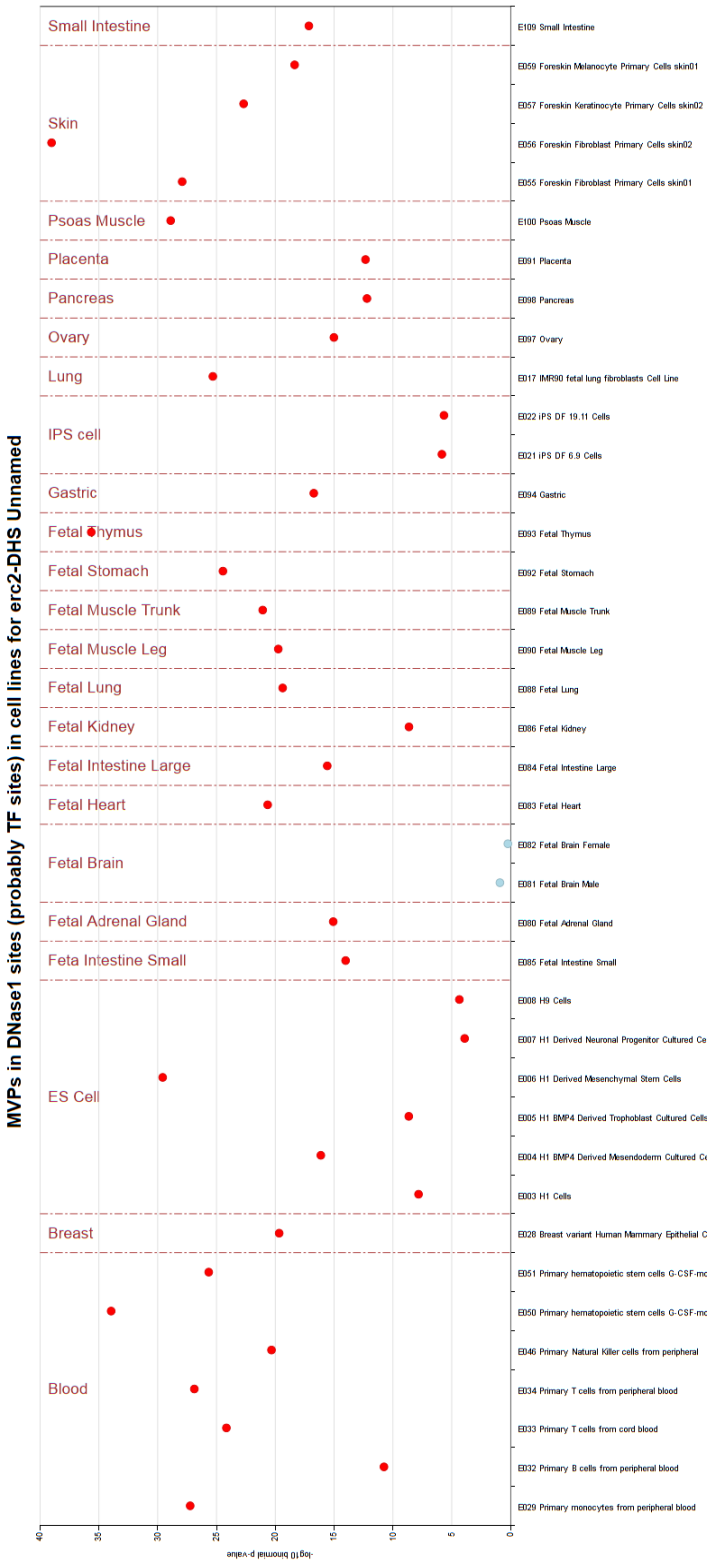
pT	0.0001	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4	0.5	1
# SNPs	805	2380	5375	7868	18981	27644	40232	49621	56826	62566	77651

Supplementary figure 2-6. Effect of processing 22q11.2 dataset without CLOZUK data and effect of sample origin. The p-value represented on top of bars represent the association of PRS with the phenotype in the model for each p-value threshold (shown by colours, cf. legend on top). **(A) Proportion of variance in schizophrenia explained by polygenic risk score (PRS) for 22q11.2 samples processed without CLOZUK data.** The same 22q11.2 samples analysed with CLOZUK data previously were processed independently (N=105 22q11.2 samples with schizophrenia and N= 171 22q11.2 samples without schizophrenia) and SNPs (N = 4003865 SNPs, around 20 times more than in the co-analysis with CLOZUK data due to the relatively modest overlap of quality SNPs between both datasets). The data processing, quality control steps and analysis were similar to what is detailed in the methods, without merging the data with the CLOZUK sample. **(B) Proportion of variance in schizophrenia explained by polygenic risk score (PRS) for 22q11.2 samples processed without CLOZUK data, after including the sample origin as a covariate in the regression model. (C) Number of independent SNPs included in the analysis for each p-value threshold pT.**

Supplementary information: Effect of the 22q11.2 deletion on DNA methylation

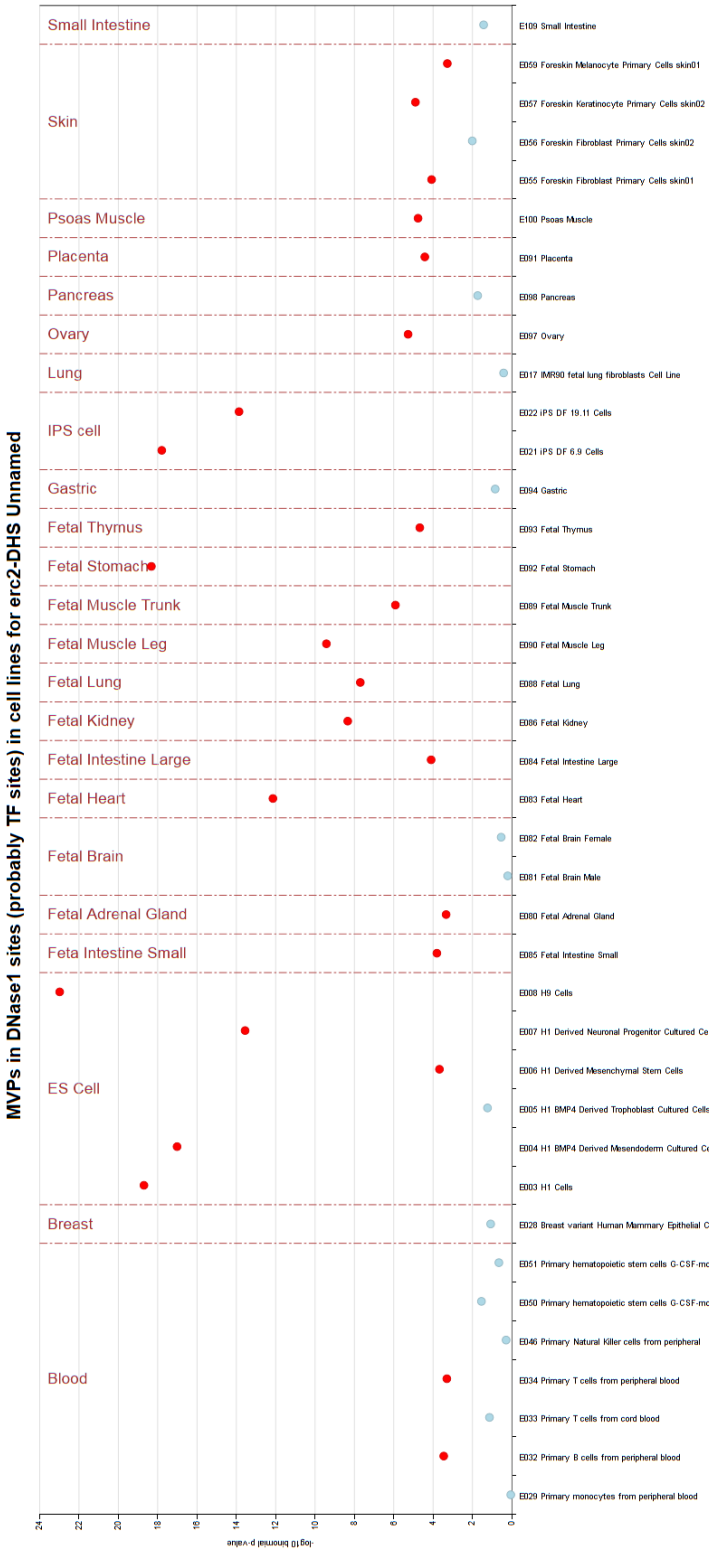


Supplementary figure 2-1. Density plots of methylation levels for each sample before (A) or after (B) ssNoob normalisation. Control samples are represented in green, 22q11.2DS cases in orange. Plots generated with the *ChAMP* R package.

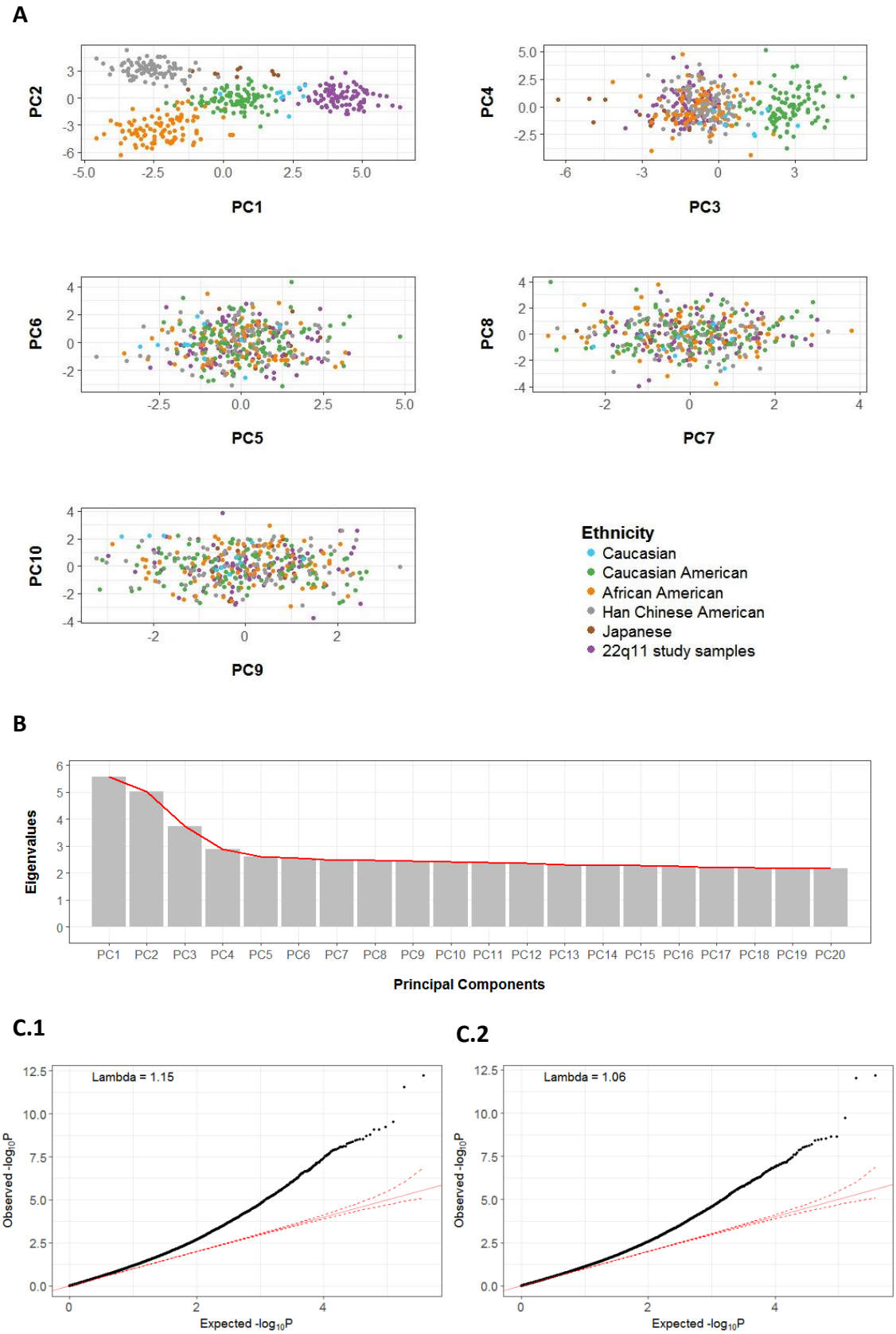


A

Supplementary figure 3-2. eFORGE analysis of enrichment of tissue-specific signal of the top 1000 most significant probes (A) Before correction for cell-type composition **(B, next page)** After Correction for cell-type composition. Red dots represent probes cluster for which FRD q-value < 0.01. Plots generated online on the eForge website.

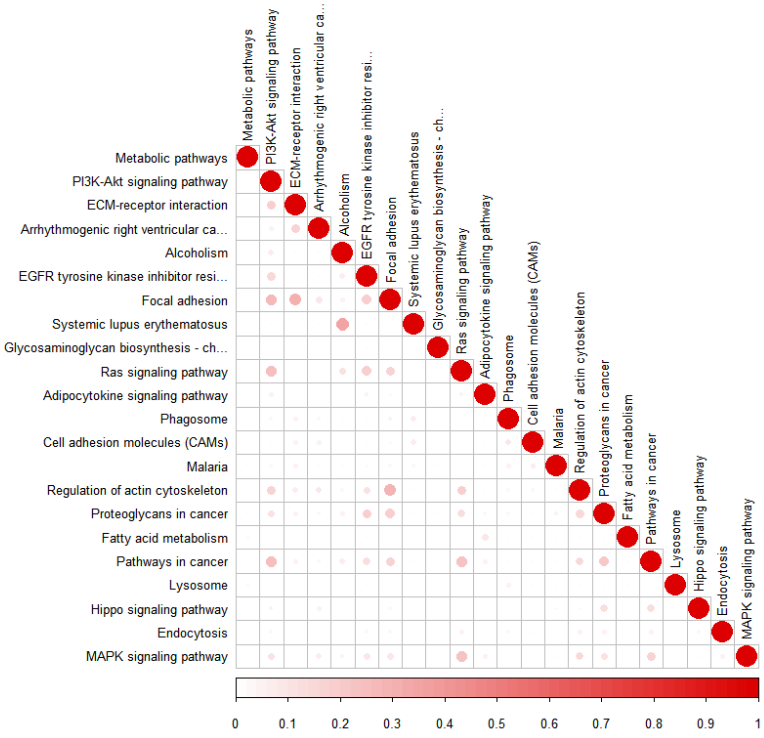

MVPs in DNase1 sites (probably TF sites) in cell lines for erc2-DHS Unnamed
Small Intestine
E109 Small Intestine
Skin
E059 Foreskin Melanocyte Primary Cells skin01
E057 Foreskin Keratinocyte Primary Cells skin02
E056 Foreskin Fibroblast Primary Cells skin02
E055 Foreskin Fibroblast Primary Cells skin01
Psoas Muscle
E100 Psoas Muscle
Placenta
E091 Placenta
Pancreas
E098 Pancreas
Ovary
E097 Ovary
Lung
E017 IMR90 fetal lung fibroblasts Cell Line
IPS cell
E022 iPS DF 19.11 Cells
E021 iPS DF 6.9 Cells
Gastric
E094 Gastric
Fetal Thymus
E093 Fetal Thymus
Fetal Stomach
E092 Fetal Stomach
Fetal Muscle Trunk
E089 Fetal Muscle Trunk
Fetal Muscle Leg
E090 Fetal Muscle Leg
Fetal Lung
E088 Fetal Lung
Fetal Kidney
E086 Fetal Kidney
Fetal Intestine Large
E084 Fetal Intestine Large
Fetal Heart
E083 Fetal Heart
Fetal Brain
E082 Fetal Brain Female
E081 Fetal Brain Male
Fetal Adrenal Gland
E080 Fetal Adrenal Gland
Feta Intestine Small
E085 Fetal Intestine Small
ES Cell
E008 H9 Cells
E007 H1 Derived Neuronal Progenitor Cultured Ce
E006 H1 Derived Mesenchymal Stem Cells
E005 H1 BMP4 Derived Trophoblast Cultured Cells
E004 H1 BMP4 Derived Mesendoderm Cultured Ce
E003 H1 Cells
Breast
E028 Breast variant Human Mammary Epithelial C
Blood
E051 Primary hematopoietic stem cells G-CSF-mc
E050 Primary hematopoietic stem cells G-CSF-mc
E046 Primary Natural Killer cells from peripheral
E034 Primary T cells from peripheral blood
E033 Primary T cells from cord blood
E032 Primary B cells from peripheral blood
E029 Primary monocytes from peripheral blood
24
22
20
18
16
14
12
10
8
6
4
2
0
-log10 binomial p-value

B

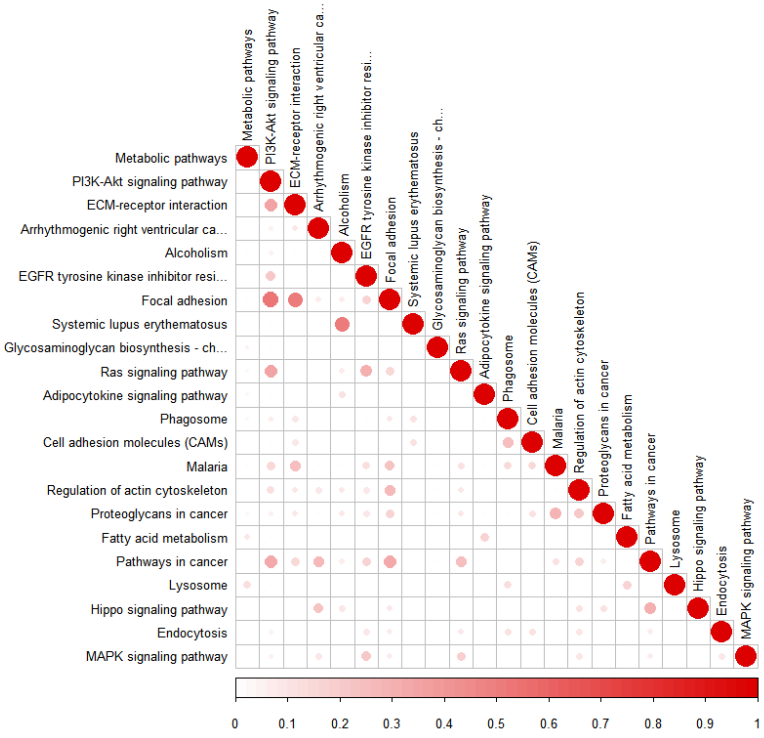


Supplementary figure 3-3. Population stratification in DNA-methylation samples (A) Principal component analysis discrimination of samples from different ethnicity. **(B)** Scree plot (variance explained by principal components (PC)). **(C)** Correction of the genomic inflation due to population stratification (**C.1**/**C.2**: before/after correction). Lambda = genomic inflation factor. A lambda close to 1 indicate that results are not inflated by unaccounted variance. The red line represents the uniform distribution; red dotted lines represent the 95% confidence interval.

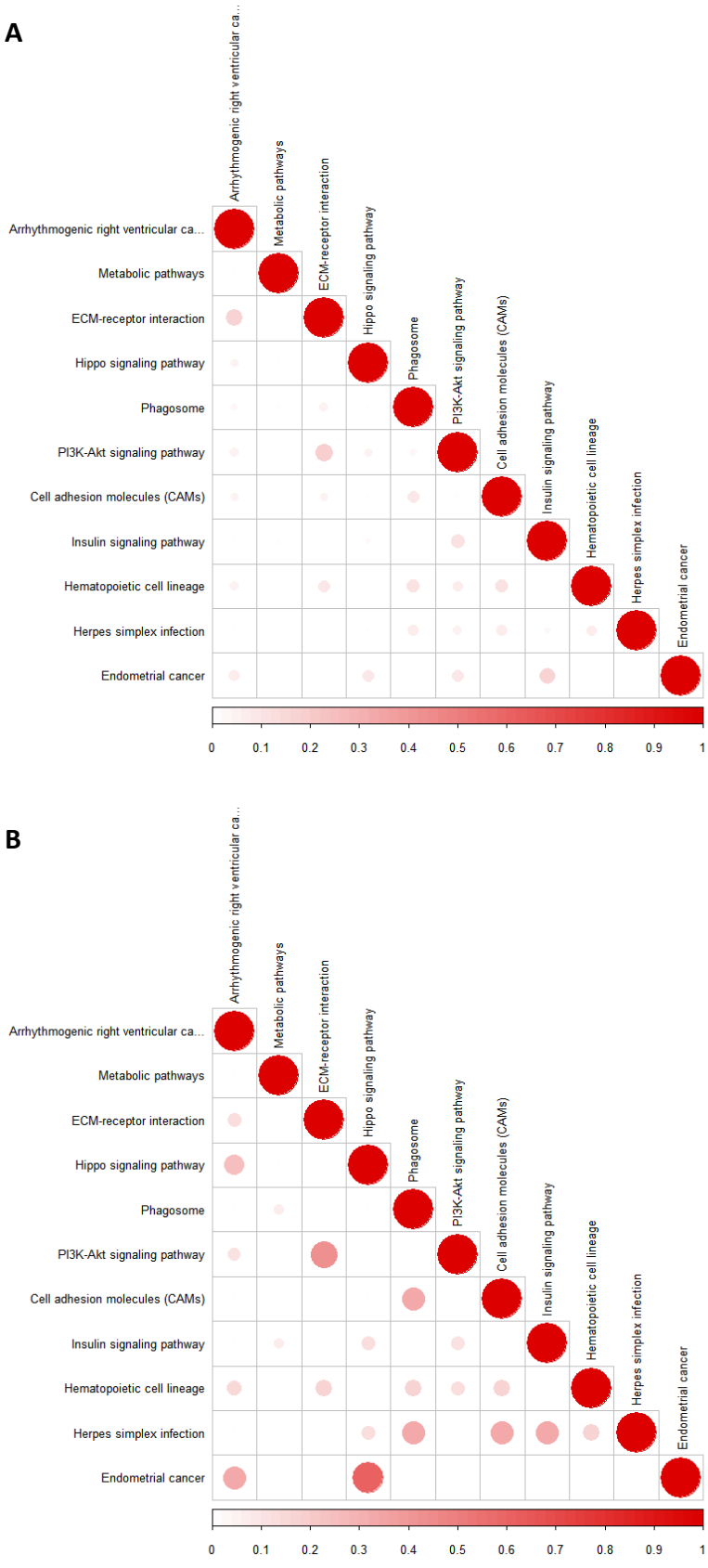
A



B



Supplementary figure 3-4. Overlap between enriched KEGG pathways (for DMPs FDR < 0.05, |Beta| > 0), including all genes within the pathways (A) or only genes within the pathways that contain at least one DMP (B). The size and color of dots represent the overlap between pathways. Values of 0 and 1 represent a 0% or 100% overlap respectively between the two pathways. DMP: differentially methylated probe. Plots generated with the *corrplot* R package.



Supplementary figure 3-5. Overlap between enriched KEGG pathways (for DMPs FDR < 0.05, |Beta| > 0.05), including all genes within the pathways (A) or only genes within the pathways that contain at least one DMP (B). The size and color of dots represent the overlap between pathways. Values of 0 and 1 represent a 0% or 100% overlap respectively between the two pathways. DMP: differentially methylated probe. Plots generated with the *corrplot* R package.

A.1

Pathway	N	DE	FDR
Metabolic pathways	1186	11	3.5e-02

A.2

Pathway	N	DE	FDR
Phagosome	139	4	3.3e-02

B.1

Pathway	N	DE	FDR
Alcoholism	163	6	2.1e-04
Focal adhesion	190	7	2.1e-04
PI3K-Akt signaling pathway	317	8	2.1e-04
Metabolic pathways	1186	12	3.6e-04
Systemic lupus erythematosus	116	4	2.5e-03
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	69	4	3.4e-03
ECM-receptor interaction	79	4	4.0e-03
Pathways in cancer	383	7	4.2e-03
Adipocytokine signaling pathway	66	3	2.2e-02
Hippo signaling pathway	151	4	3.6e-02
Mannose type O-glycan biosynthesis	22	2	3.6e-02
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	19	2	4.4e-02
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	23	2	4.4e-02
EGFR tyrosine kinase inhibitor resistance	77	3	4.9e-02

B.2

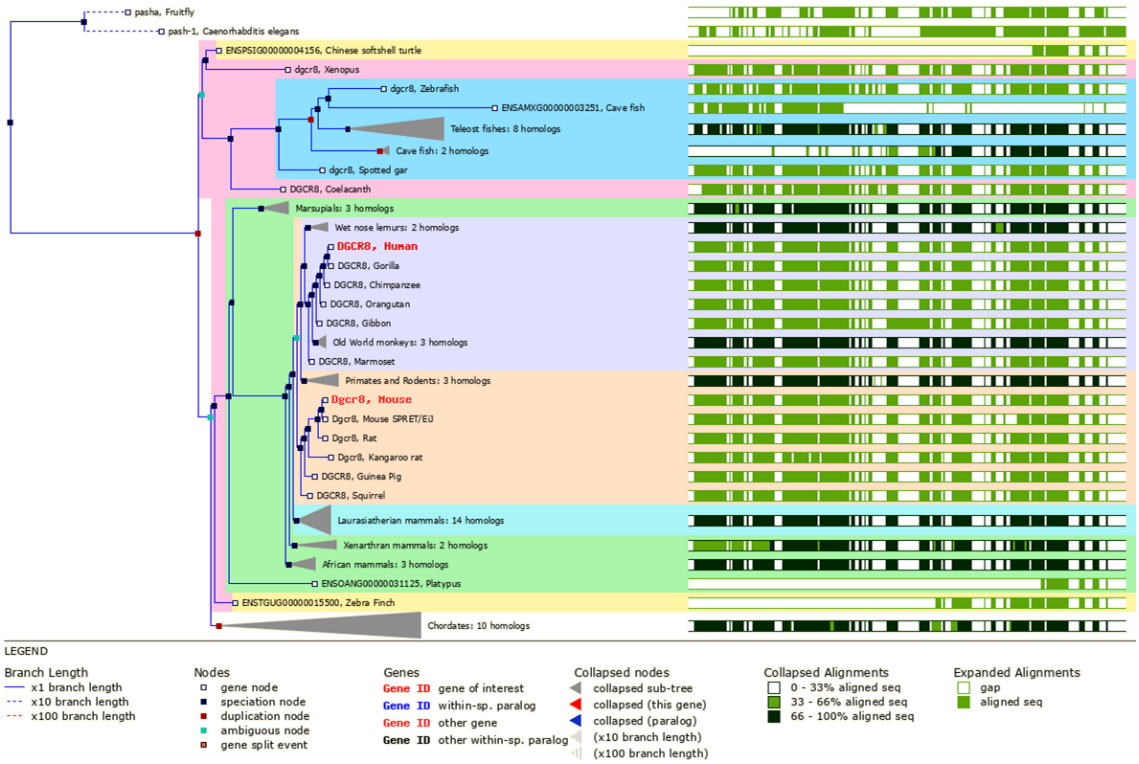
Pathway	N	DE	FDR
Hippo signaling pathway	151	4	1.0e-02
Metabolic pathways	1186	7	1.0e-02
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	69	3	1.4e-02
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	23	2	2.9e-02
Insulin signaling pathway	131	3	3.1e-02

Supplementary figure 3-6. KEGG pathway enrichment analysis for probes hypermethylated (A) and hypomethylated (B) in 22q11.2DS samples. (A.1) DMPs with FDR < 0.05 and ΔBeta > 0 (N_{DMP} = 380). **(A.2)** DMPs with FDR < 0.05 and ΔBeta > 0.05 (N_{DMP} = 246). **(B.1)** DMPs with FDR < 0.05 and ΔBeta < 0 (N_{DMP} =281). **(B.2)** DMPs with FDR < 0.05 and ΔBeta < 0.05 (N_{DMP} = 155). N: number of gene in the pathway; N_{DMP}: number of DMP tested; DE: number of genes associated with at least one DMP; FDR: false discovery rate (the number of probes per gene is taken into account to adjust the probability of significant differential expression). Only significant KEGG pathway enrichment (FDR < 0.05) are shown.

Location	Width	# CpGs	Stouffer	deltaBeta	OverlappingGenes	NearestGenes	BloodBrain
chr20:32856747-32857227	481	6	4.3e-08	0.11	, ASIP	ASIP	TRUE
chr1:23003462-23003692	231	3	8.8e-08	0.05	NA	C1QB	FALSE
chr5:66461884-66462662	779	4	6.9e-06	-0.09	MAST4	MAST4	FALSE
chr17:75789279-75789529	251	5	7.1e-05	-0.08	NA	NA	FALSE
chr10:101282185-101283090	906	7	3.7e-04	-0.09	NA	NA	TRUE
chr7:25893245-25894005	761	5	1.6e-03	-0.07	NA	MIR148A	FALSE
chr2:164204628-164205343	716	7	3.8e-03	-0.06	NA	FIGN	FALSE
chr8:21908746-21909033	288	2	9.6e-03	-0.03	DMTN	DMTN	FALSE
chr4:39448432-39448975	544	4	1.0e-02	0.09	KLB	KLB	TRUE
chr6:28921733-28922226	494	6	1.1e-02	0.05	NA	GPX6	FALSE
chr3:194014481-194015171	691	7	1.2e-02	-0.06	LINC00887	NA	FALSE
chr14:95330880-95330984	105	2	1.8e-02	-0.12	NA	NA	TRUE
chr13:50706583-50707065	483	6	2.2e-02	-0.02	DLEU1	DLEU1	FALSE
chr7:55073022-55073248	227	2	3.6e-02	-0.12	NA	EGFR	FALSE

Supplementary figure 3-7. Differentially methylated regions (DMRs) not associated with a gene promoter region. DeltaBeta: average beta value differences for the probes (deltaBeta > 0: hypermethylation in 22q11.2DS samples); OverlappingGenes: genes overlapping the DMR; NearestGenes: Gene closest to the DMR Location: genomic location. Stouffer: combined false discovery rates for the probes, with the Stouffer method; Width: DMR size.

Supplementary information: CRISPR/Cas9 genome editing of DGCR8 in human embryonic stem cells



Supplementary figure 4-1. DGCR8 orthologues gene tree. Source: ensemble.org.

Figure	Primer set	Forward primer	Reverse primer	Ref.
Figure 1-4	A	GAGTGGATTGCTGTGCTCTG	AGACTGCCTTGGGAAAAGC	
	B	GCCTTCTGTGTGTTCCAGAAA	CTAAAGCGCATGCTCCAGAC	
	C	TGGGCTCTATGGCTTCTGAG	GCAAAACACGCTGTTCAGAC	
	D	GGGCTCGAGATCCACTAGTTC	CTTCTTCCAGCAGAGCATCC	
	E	TGCAGAGGTAATGGACGTTG	AGCTCTCGGTAAAGCTCACG	
Figure 4-6, ***Figure 4-8***, ***Figure 4-9***	DGCR8 Exon 1/2	CTTTCCCGGCTGTGGTTTGG	CTGGCGACTAAGCGAGTCTG	
Figure 4-6, ***Figure 4-9***	DGCR8 miRNA sites	GGCAGTGGTTCTAAAAGCTGTC	AGGTGGTGGAGACTGCTCAC	
	DGCR8 Target site	CGGAGCTTCTCTTCTCTCCA	ACTCCTGCAGCTCTCGGTAA	
	DGCR8 Exon 12/14	CAAGCAGGAGACATCGGACAAG	CACAATGGACATCTTGGGCTTC	481
Figure 4-7, ***Figure 4-9***	pri-miR-185	AGACCTGCTGGCTAGAGCTG	CAAGGGAAGGCCATAAACAG	
	pri-let-7a-1	GATTCCTTTTCACCATTCACCCTGGATGTT	TTTCTATCAGACCGCCTGGATGCAGACTTT	415
	pri-miR-16-1	GAAAAGGTGCAGGCCATATTGT	CGCCAATATTTACGTGCTGCTA	
	pri-miR-17-92a	GGGAAACTCAAACCCCTTTCTAC	CAACAGGCCGGGACAAGT	488
	pri-miR-17-19a	TGCCCTAAGTGCTCCTTCTG	AAATAGCAGGCCACCATCAG	
	pri-miR-20-19a	GCCCAATCAAACTGTCCTGT	ACAATCCCCACCAAACTCAA	
Figure 4-8, ***Figure 4-9***	Nestin	GTGCAGAGGTGGGAAGATACG	CCTGCTCTACCACCTCCTCCT	
	Tuj1	CATGGACAGTGTCCGCTCAG	CAGGCAGTCGCAGTTTTCAC	
	Pax6	AATAACCTGCCTATGCAACCC	AACTTGAACTGGAACTGACACAC	
	Neurog2	TCAGACATGGACTATTGGCAG	GGGACAGGAAAGGGAACC	425
	NeuroD1	ACGACCTCGAAACCATGAAC	CTTCCAGGTCCTCATCTTCG	
	GAPDH	ACGACCCCTTCATTGACCTCAACT	ATATTTCTCGTGGTTCACACCCAT	

Supplementary table 4-1. Table of primer sequences.

Supplementary information: Transcriptome analysis of DGCR8 knock-out lines

A

GO term	GO type	Adjusted p-value
cell morphogenesis involved in neuron differentiation	BP	5.6E-05
axon development	BP	5.6E-05
locomotion	BP	4.9E-04
cell adhesion	BP	5.2E-04
biological adhesion	BP	5.2E-04
movement of cell or subcellular component	BP	5.2E-04
regulation of anatomical structure morphogenesis	BP	5.2E-04
developmental growth involved in morphogenesis	BP	5.2E-04
extracellular matrix	CC	6.3E-04
proteinaceous extracellular matrix	CC	6.3E-04
tissue development	BP	6.8E-04
morphogenesis of a branching structure	BP	6.8E-04
regulation of cell projection organization	BP	1.6E-03
regulation of cell morphogenesis involved in differentiation	BP	1.7E-03
response to external stimulus	BP	1.8E-03
regulation of developmental growth	BP	1.9E-03
glycosaminoglycan binding	MF	1.9E-03
calcium ion binding	MF	1.9E-03
sulfur compound binding	MF	1.9E-03
heparin binding	MF	1.9E-03
tissue morphogenesis	BP	3.7E-03
enzyme linked receptor protein signaling pathway	BP	4.2E-03
chemotaxis	BP	5.4E-03
taxis	BP	5.4E-03
regulation of neuron projection development	BP	5.4E-03
system process	BP	9.5E-03
head development	BP	9.5E-03
regulation of cell proliferation	BP	1.1E-02
negative regulation of response to stimulus	BP	1.3E-02
regulation of growth	BP	1.4E-02
cell motility	BP	1.4E-02
localization of cell	BP	1.4E-02
growth	BP	1.5E-02
receptor binding	MF	1.6E-02
signaling receptor activity	MF	1.7E-02
sodium ion homeostasis	BP	1.7E-02
response to axon injury	BP	1.8E-02
regulation of response to wounding	BP	2.0E-02
positive regulation of peptidyl-serine phosphorylation	BP	2.3E-02
positive regulation of developmental process	BP	2.3E-02
response to lipid	BP	2.7E-02
cell proliferation	BP	2.8E-02
negative regulation of multicellular organismal process	BP	2.8E-02
response to nicotine	BP	2.8E-02
hindlimb morphogenesis	BP	2.8E-02
single-organism behavior	BP	3.2E-02
regulation of extent of cell growth	BP	3.4E-02
positive regulation of kinase activity	BP	3.5E-02
regulation of body fluid levels	BP	3.5E-02
regulation of vascular permeability	BP	3.5E-02
regulation of cardiac muscle tissue development	BP	3.5E-02
negative regulation of developmental process	BP	4.1E-02
cardiac muscle tissue development	BP	4.1E-02
negative regulation of cell adhesion	BP	4.4E-02
peptidase inhibitor activity	MF	4.6E-02
protein tyrosine kinase activity	MF	4.6E-02
signal transducer activity	MF	4.6E-02
long-term synaptic potentiation	BP	4.7E-02

Supplementary table 5-1. Gene Ontology gene-set enrichment analysis of high-confidence genes differentially expressed in *DGCR8*$^{+/-}$ and 22q11.2DS neural progenitor cells. Legend on next page.

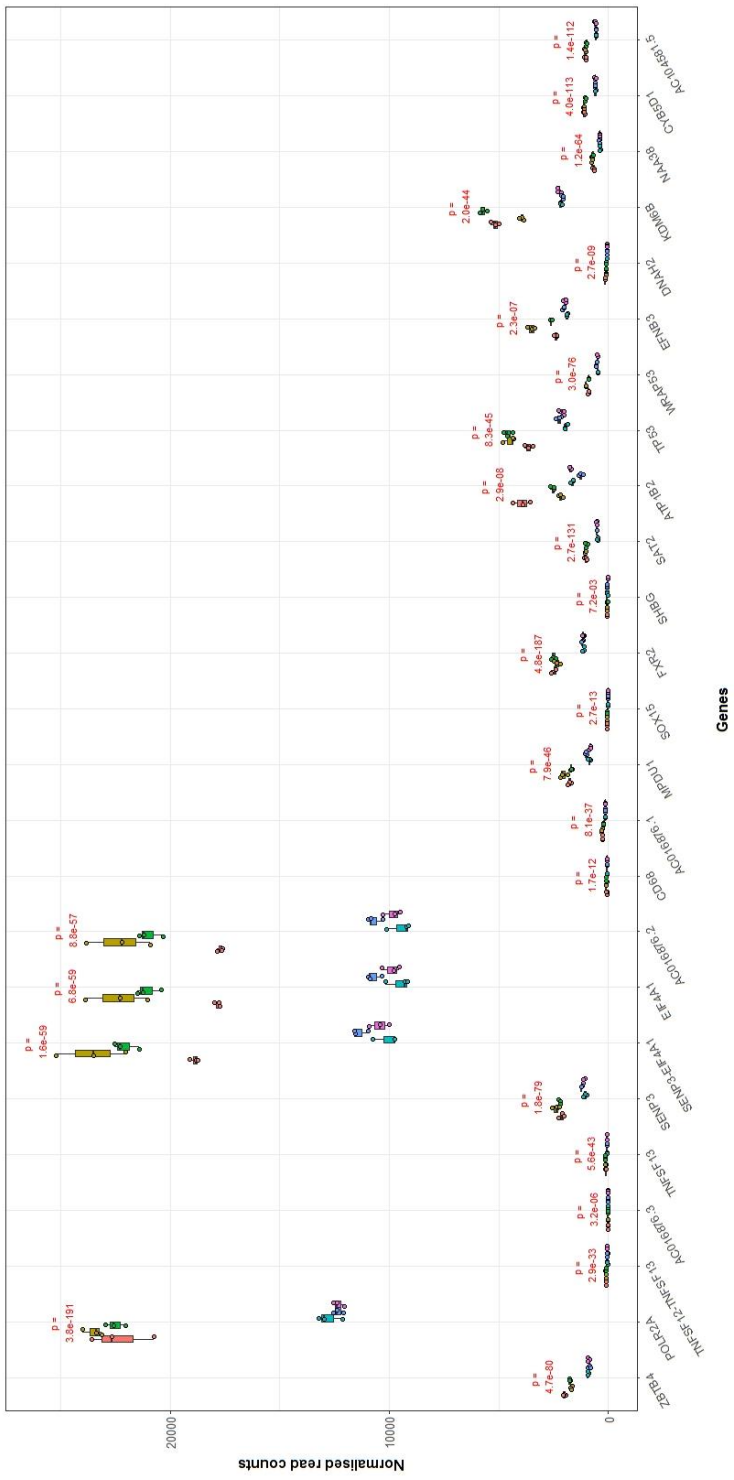
B

GO term	GO type	Adjusted p-value
calcium ion binding	MF	2.6E-06
cell-cell adhesion via plasma-membrane adhesion molecules	BP	7.2E-06
cell adhesion	BP	7.2E-06
biological adhesion	BP	7.2E-06
proteinaceous extracellular matrix	CC	2.8E-02
extracellular space	CC	2.8E-02

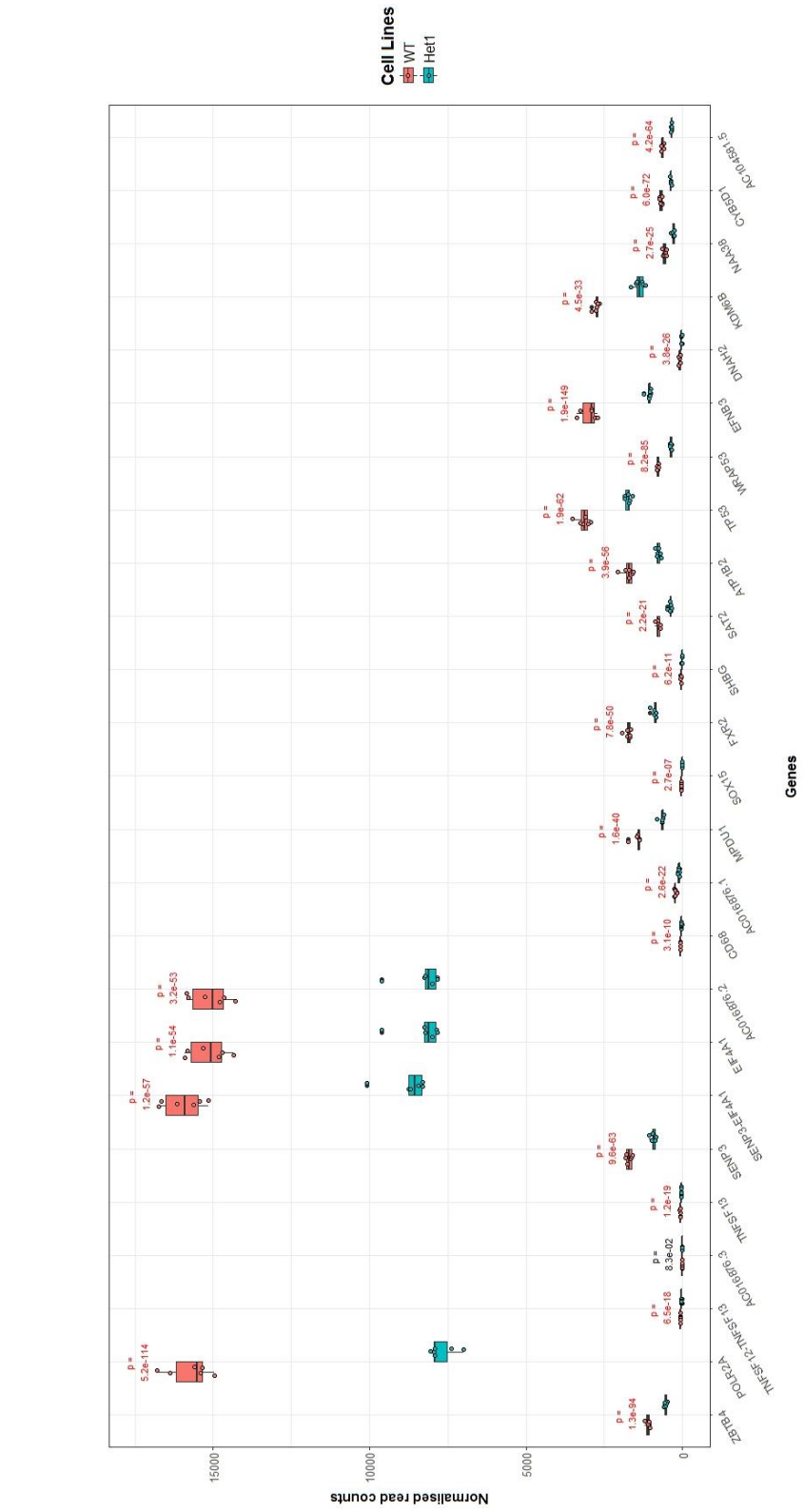
C

GO term	GO type	Adjusted p-value
neuron differentiation	BP	1.4E-05
head development	BP	4.6E-05
locomotion	BP	4.4E-04
cell projection organization	BP	5.8E-04
movement of cell or subcellular component	BP	3.1E-03
heparin binding	MF	3.5E-03
regulation of nervous system development	BP	3.8E-03
regulation of neuron projection development	BP	3.8E-03
glycosaminoglycan binding	MF	4.2E-03
sulfur compound binding	MF	5.3E-03
response to nicotine	BP	7.7E-03
regulation of developmental growth	BP	7.8E-03
regulation of cell morphogenesis involved in differentiation	BP	8.7E-03
regulation of anatomical structure morphogenesis	BP	9.7E-03
animal organ morphogenesis	BP	1.1E-02
mitotic spindle organization	BP	1.5E-02
transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	MF	1.9E-02
cell motility	BP	1.9E-02
localization of cell	BP	1.9E-02
developmental growth involved in morphogenesis	BP	1.9E-02
growth	BP	2.3E-02
sodium ion homeostasis	BP	2.3E-02
regulation of growth	BP	2.4E-02
positive regulation of developmental process	BP	2.4E-02
epithelium development	BP	2.6E-02

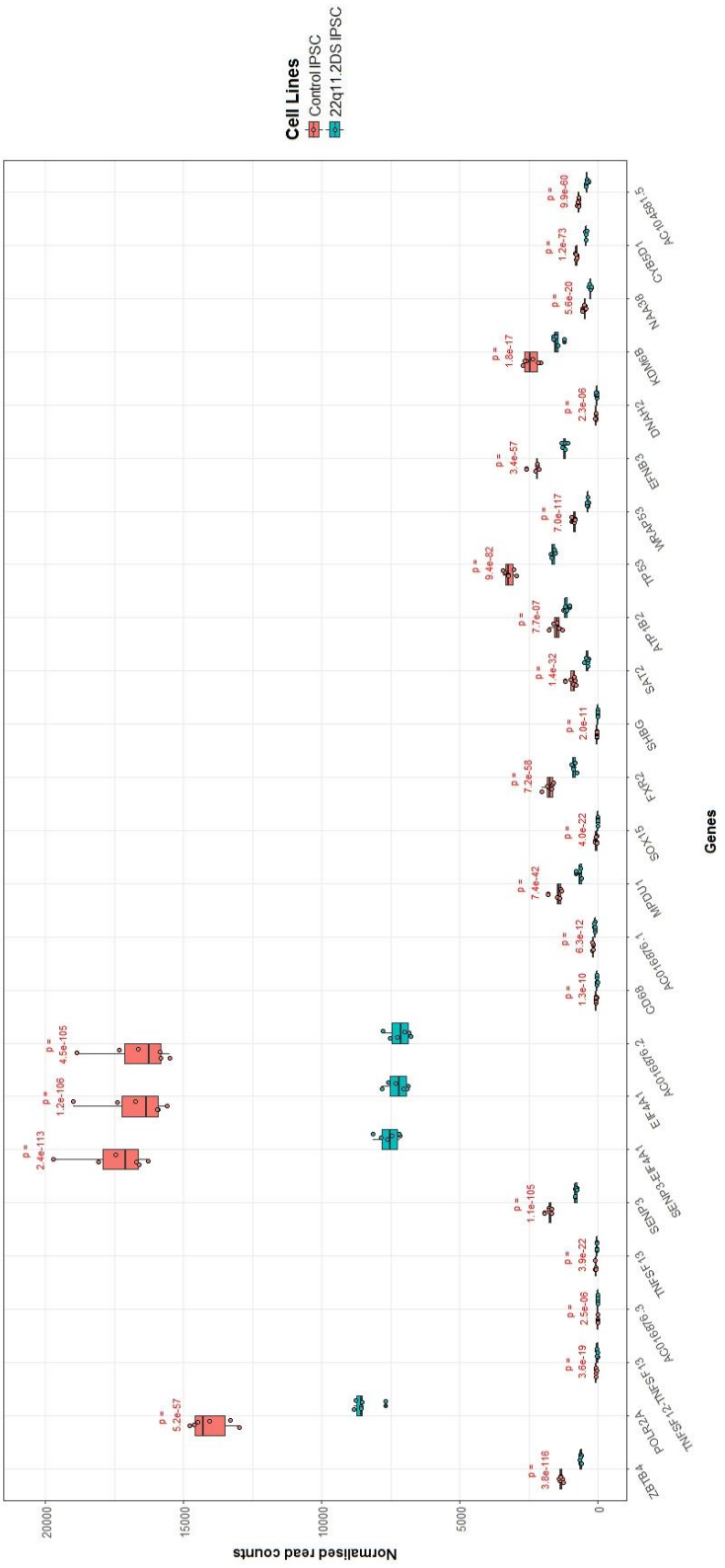
Supplementary table 5-1. Gene Ontology gene-set enrichment analysis of high-confidence genes differentially expressed in *DGCR8*$^{+/-}$ and 22q11.2DS neural progenitor cells. (**A**) All significant genes. (**B**) Upregulated genes only. (**C**) Downregulated genes only. Differentially expressed genes are defined as genes with an FDR corrected p-value < 0.01. Significantly enriched terms have an adjusted p-value < 0.05 (Benjamini-Hochberg procedure). BP: biological process; CC: cellular component; MF: molecular function.



Supplementary figure 5-1. Difference in expression for significant genes between DGCR8 KO lines, 22q11.2 DS line and their respective controls at the locus 17p13.1. Genes are ordered by position on chromosome 17. (**A**) Experiment 1. (B) Experiment 2, DGCR8$^{+/-}$ Het1 line versus WT line. (C) Experiment 2, 22q11.2 DS hiPSC line versus control IPSC line. The adjusted p-value for each gene is represented. Significant p-values are shown in red.



Supplementary figure 5-1. Difference in expression for significant genes between DGCR8 KO lines, 22q11.2 DS line and their respective controls at the locus 17p13.1. Genes are ordered by position on chromosome 17. (cont.)



C

Supplementary figure 5-1. Difference in expression for significant genes between DGCR8 KO lines, 22q11.2 DS line and their respective controls at the locus 17p13.1. Genes are ordered by position on chromosome 17. (cont.)