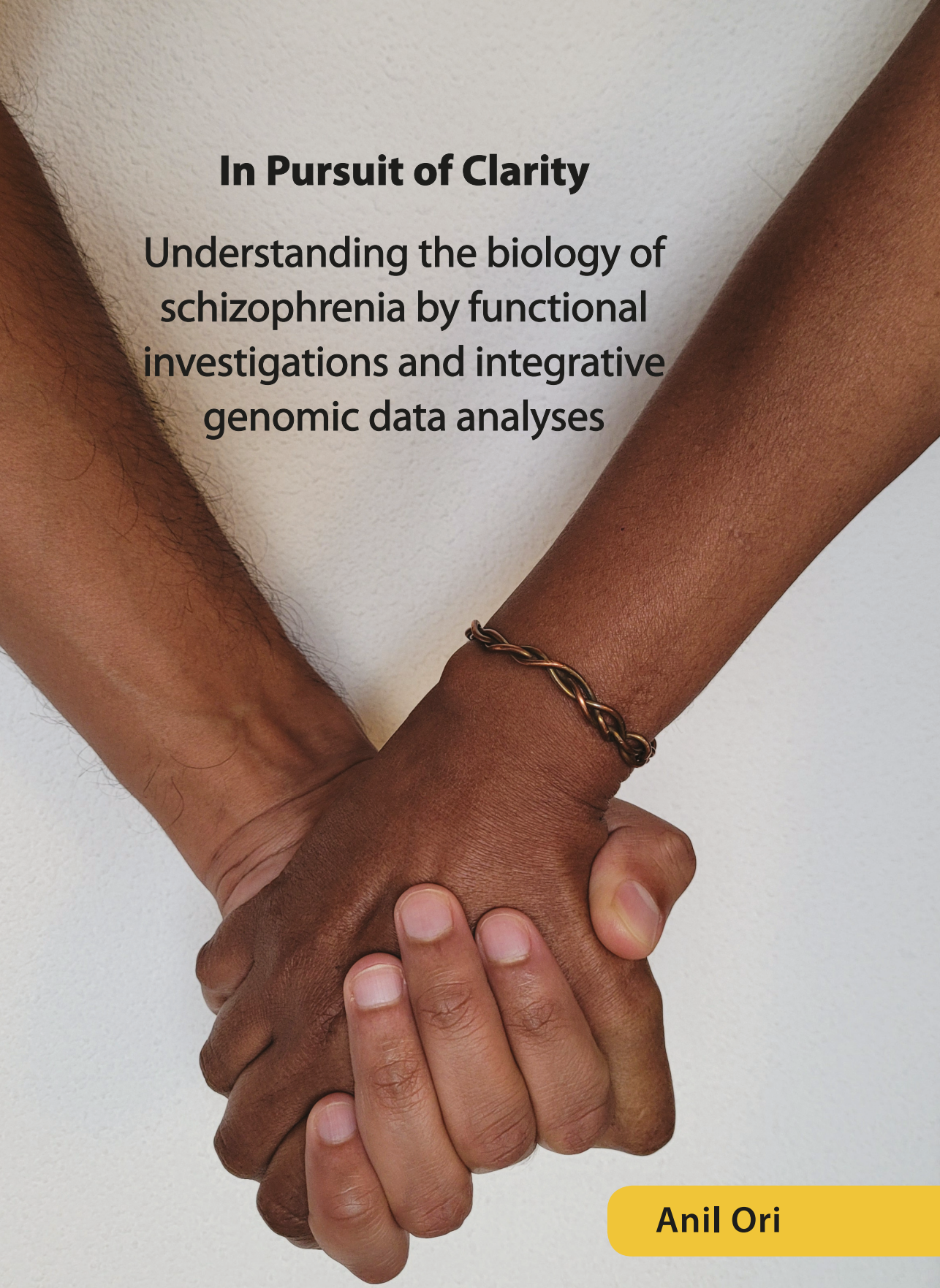
In Pursuit of Clarity

Understanding the biology of schizophrenia by functional investigations and integrative genomic data analyses



Anil Ori

In Pursuit of Clarity

Understanding the biology of schizophrenia by functional investigations and integrative genomic data analyses

De zoektocht naar verheldering

Het begrijpen van de biologie van schizofrenie met behulp van functioneel onderzoek en integratieve genomische data-analyse

Anil Ori

Printed by Ipskamp Printing
Enschede, the Netherlands

Design & layout Bianca Pijl, www.pijlldesign.nl
Groningen, the Netherlands

ISBN 978-94-6421-634-9

Proofreading of introduction and discussion text: Samuel Ha, Groningen, the Netherlands

In Pursuit of Clarity

Understanding the biology of schizophrenia
by functional investigations and integrative
genomic data analyses

De zoektocht naar verheldering

Het begrijpen van de biologie van schizofrenie
met behulp van functioneel onderzoek en
integratieve genomische data-analyse

Thesis

To obtain the degree of Doctor from the
Erasmus University Rotterdam
By command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defense shall be held on

Tuesday February 15, 2022, at 15:30h

by

Anil Pravin Surendredath Ori
born in Paramaribo, Suriname

Erasmus University Rotterdam

Erasmus

Doctoral Committee

Promotors
Prof. dr. ir. R.A. Ophoff
Prof. dr. S.A. Kushner

Other Members
Prof. dr. M.A. Ikram
Prof. dr. D. Posthuma
Prof. dr. B.W.J.H. Penninx

Paranymphs
Annabel Vreeker
Rafida Abdoelrahman

Table of contents

Preface

This Ph.D. dissertation is a collection of my research on the genomic causes and consequences of schizophrenia. My pursuit in research on the illness stems both from my interest in functioning of the brain and my personal lived experience. At age 45, my mother experienced her first psychosis, which was later linked to a clinical diagnosis of schizophrenia. As her journey of treatment and recovery from the psychosis started, so did my journey of understanding the illness and supporting her in the most loving way possible. As a family, the illness was unbeknown to us and in times of distress and confusion that followed, I sought answers.

In pursuit of clarity, I started my Biomedical Sciences Master's research program at the Utrecht University, the Netherlands. I enrolled in the Biology of Disease research track. I was on a mission to learn about disease mechanisms, in particular that of schizophrenia. During my studies, I worked as a research intern at the Department of Human Genetics, University Medical Center Utrecht in the Netherlands, on genome-wide gene expression signatures in human post-mortem brain tissue of people diagnosed with schizophrenia. Soon after successful completion of my master's studies, I moved to the United States of America, to work at the Center for Neurobehavioral Genetics at University of California, Los Angeles (UCLA), as a research scientist. Here, I was surrounded by international leaders in human genetic and psychiatric research. In the research group of Prof. Dr. Roel A. Ophoff, I further shaped my training in psychiatric genomic science and conducted the research studies on the biology of schizophrenia as described in this doctoral thesis book.

On a personal level, my aim was to learn about the illness and deepen my understanding of its phenotypic and biological characteristics and complexity. Having loved ones who have been diagnosed with the illness, I have seen the impact on and suffering of those affected up close. In gaining understanding of the illness, I hope to gain understanding of their journey as well. On a scientific level, I aimed to go beyond the findings of large-scale genetic studies and conducted research that uses state-of-the-art methodology and integrative genomic data analyses to identify new pieces to the puzzle of schizophrenia biology in the post-GWAS era. I am grateful to the Graduate School of the Erasmus University Rotterdam and Erasmus Medical Center for accepting me into their graduate program as an external Ph.D. candidate.

Schizophrenia is a severe illness with a heartbreaking impact on those who suffer and their families. Clarifying the underlying biological mechanisms of schizophrenia is a necessary step for the development of more effective treatments and further humanization of the illness. I hope my research will contribute towards achieving this goal.

Voorwoord

Dit proefschrift is een verzameling van mijn doctoraal onderzoek naar de genomische oorzaken en gevolgen van schizofrenie. Mijn drijfveer om onderzoek te doen naar deze ziekte komt voort uit mijn interesse naar het menselijk brein en uit mijn geleefde ervaring. Op 45-jarige leeftijd heeft mijn moeder een psychose ervaren wat later gelinkt werd aan een klinische diagnose van schizofrenie. Toen haar traject van behandeling en herstel van de psychose begon, begon ook mijn traject van het willen begrijpen van de ziekte en haar steunen op de meest liefdevolle manier als dat ik kon. We wisten weinig over de ziekte, maar zaten wel met vragen. In de tijd die volgde, ging ik opzoek naar antwoorden.

In mijn zoektocht naar verheldering, ben ik de masteropleiding Biomedische Wetenschappen met een specialisatie in ziektemechanismen gaan volgen aan de Universiteit Utrecht. Ik was vastberaden meer te leren over de biologie die ten grondslag ligt aan ziekten, vooral die van schizofrenie. Tijdens mijn studie, heb ik als onderzoekstagiaire gewerkt bij de afdeling Humane Genetica van het Universitair Medisch Centrum Utrecht waar ik onderzoek heb uitgevoerd naar genexpressie signalen in post-mortem hersenweefsel van mensen die schizofrenie hebben ervaren. Na het succesvol afronden van mijn masteropleiding, ben ik verhuisd naar de Verenigde Staten om te werken bij de Center for Neurobehavioral Genetics aan de University of California, Los Angeles (UCLA) als wetenschappelijk onderzoeker. Hier ben ik omringd geweest door vooraanstaande leiders op het gebied van humaan genetisch en psychiatrisch onderzoek. In de onderzoeksgroep van Dr. Roel Ophoff heb ik mijn training en kennis over de psychiatrische genetica verder vormgegeven en ook het onderzoek uitgevoerd die in dit proefschrift beschreven staat.

Op het persoonlijk vlak, was mijn doel om meer te leren over schizofrenie, hoe het zich tot uiting brengt en hoe diens biologie precies in elkaar steekt. Door de ervaring van mijn moeder, heb ik van dichtbij meegemaakt wat de impact van de ziekte is. Met het beter begrijpen van de ziekte, hoop ik mij ook beter te kunnen inleven in de ervaring van haar en anderen die door schizofrenie ziek zjn geworden. Op het wetenschappelijk vlak, was mijn doel om de bevindingen van grootschalige genetisch studies een stap verder te brengen door onderzoek uit te voeren dat gebruik maakt van innovatieve methodiek en integratieve genomische analyses om tot nieuwe inzichten over de ziekte te komen. Ik ben dankbaar dat de Graduate School van de Erasmus Universiteit Rotterdam en het Erasmus Medisch Centrum mij hebben toegelaten tot hun PhD programma als externe PhD kandidaat.

Schizofrenie is een ziekte die een enorme impact heeft op patiënten en hun naasten. Het beter begrijpen van de biologische ziektemechanismen die ten grondslag liggen aan schizofrenie is een noodzakelijke stap naar de ontwikkeling van meer effectievere behandeling en naar het vermenselijken van een ziekte die zo gestigmatiseerd is. Ik hoop dat mijn onderzoek hieraan zal bij dragen.

INTRODUCTION

CHAPTER 1

Outline of thesis

Human genomics is a rapidly evolving area of research that is revolutionizing our understanding of biology and our ability to improve human health. The Human Genome Project was completed in 2003, which characterized the make-up of human DNA for the first time, costing an estimated $2.7 billion. Today, sequencing a whole human genome costs less than $1,000. Collecting genome-wide gene expression or DNA methylation (DNAm) data costs $250 or less per biological sample. Advancements in biotechnology alongside new scientific understanding have fueled many large-scale genomic studies across human populations the past decade. By analyzing genetic variation at millions of sites in the genome we are now able to successfully perform a genome-wide association study (GWAS) in hundreds of thousands or even millions of individuals. These investigations have provided novel and more accurate insights into the genetic architecture of human health and disease, including that of major psychiatric illnesses. Genomic research can help accelerate the identification of genes or pathways that cause psychiatric illnesses and may aid in redefining the existing psychiatric nosology from descriptive-based assessments to more biologically driven diagnoses. As genomic technologies have become the norm in research and derived methodology, and scientific insights are clarifying our understanding of disease mechanisms, the opportunities to translate these genomic findings to the clinic are promising.

My thesis focuses on understanding the molecular causes and consequences of schizophrenia from a genomic perspective. Schizophrenia is a severe psychiatric disorder characterized by diverse psychopathology that affects almost 1% of the population. Patients present long-term symptoms and disabilities, high unemployment rates, and a life expectancy that is reduced by 15 years compared to the general population. A pressing need for more effective treatment exists, but the biological mechanisms that underlie the causes of schizophrenia are poorly understood. The past decade has seen significant advances in schizophrenia research, particularly in the application of genetics and genomics. Genetic research on schizophrenia has been a trailblazer in psychiatric genetics, prompting the first genome-wide association study (GWAS) of any psychiatric illness that identified over 100 regions in the genome that increase risk for the illness. This was a landmark moment in psychiatric genetics, providing overwhelming evidence of the biological origins of schizophrenia. The next key steps are to translate genetic findings into mechanistic insights and advance precision diagnostic and treatment tools. Complementary approaches that investigate disease consequences are important as well. Advancing our understanding of the molecular consequences of antipsychotic medication and its side-effects or that of environmental and lifestyle factors, for example, can help improve the quality of life of current and future patients. Schizophrenia is a complex illness that is caused by interactions between various combinations of genetic, biological, environmental, psychological, and/or social factors. This dissertation focuses on the genomic signatures of schizophrenia. Using both *in vitro* experiments and case-control cohorts, multiple genomic technologies, and state-of-the- art statistical methodology, I aim to disentangle, and if possible, clarify some of the biological complexity that underlies the illness.

My dissertation has several key aims: (1) to translate findings from GWAS into disease biology by investigating the functional mechanisms that underlie schizophrenia heritability, (2) to map the molecular profile of clozapine response by genome-wide gene expression and DNA methylation data analyses, and (3) to investigate the molecular consequences of the illness by quantifying biological age using DNAm clocks. **Part 1** of my thesis describes functional investigations of schizophrenia biology using *in vitro* experimental systems. I show how to integrate heritability measured from large-scale human population studies with molecular signatures obtained from *in vitro* model systems. Human neural stem cells (hNSCs) and lymphoblastoid cell lines (LCLs) are used to capture dimensions of early neuronal development and molecular responses to antipsychotic medication, respectively. In Chapter 3, I describe how polygenic risk of schizophrenia concentrates in a specific longitudinal gene cluster that is important for synaptic function during neuronal differentiation. In Chapter 4, I describe how clozapine exposure induces widespread changes in gene expression related to cholesterol metabolism, but not schizophrenia genetic risk. Overall, the first part of my dissertation highlights the value of combining *in vitro* experimental systems with integrative genomic data analyses and thereby the value of translating findings from the bench to human biology.

Part 2 describes an investigation of differential aging in schizophrenia using DNAm data from whole blood. DNAm profiles are modifiable by lifestyle and environmental influences and can be used to track the pace of biological aging. DNAm clocks are recently developed biomarkers of aging and represent algorithms that can estimate an individual's biological age. These clocks are predictive of current and future health, including mortality risk. In the second part of my thesis, I first provide a comprehensive evaluation of data processing and normalization strategies to optimize the performance of DNAm-based algorithms, including DNAm clocks. Then, by aggregating data across four European cohorts, I describe a meta-analysis that shows that individuals diagnosed with schizophrenia age differently than non-psychiatric control individuals for the first time. This work describes one of the largest DNAm studies on schizophrenia and highlights how specific and identifiable groups of patients, particularly women with schizophrenia in later adulthood, appear significantly older in their biological age, a phenomenon associated with an increased mortality risk. **Part 3** discusses these results and their broader implications for schizophrenia research and other psychiatric illnesses in general. I discuss the lessons learned further and outline future research and clinical implications.

In **Chapter 2**, I provide a more detailed introduction to the clinical presentation and genetic aspects of schizophrenia. I describe the presentation of the phenotype and the medical criteria used to determine a diagnosis. I further outline recent progress and results from genetic and genomic studies.

Part 1 | Functional Investigations of schizophrenia biology

In **Chapter 3**, I describe a longitudinal model of human neuronal differentiation for studying schizophrenia polygenic risk. Schizophrenia heritability has been reported to be enriched for biological processes important for early brain development and the function of neurons. Human neural stem cells were used to generate neurons across 30 days of differentiation

in an experimental laboratory model system. In this work, I describe a statistical framework for identifying longitudinal gene expression changes over time and how to integrate these molecular profiles with polygenic risk of major psychiatric illnesses. I demonstrate how schizophrenia heritability concentrates in a gene cluster important for synaptic functioning that is upregulated during neuronal differentiation. More broadly, this study shows the value of integrating genetic effect sizes estimated from large-scale genetic studies with genomic profiles identified in *in vitro* experimental systems.

In **Chapter 4**, I describe an LCL model to study clozapine biology. Clozapine is an effective antipsychotic medication that is associated with significant metabolic adverse effects, such as weight gain and in rare cases agranulocytosis. LCLs were exposed at increasing concentrations of clozapine and genome-wide gene expression and DNAm profiles investigated to gain mechanistic insights into the function of clozapine. This study demonstrates how clozapine induces widespread changes in gene expression, mainly through cholesterol and cell proliferation pathways, in contrast to more specific changes in DNA methylation levels. Gene expression changes are furthermore enriched for cholesterol genetic risk but not schizophrenia risk highlighting a potential use of LCLs to study molecular dynamics of antipsychotic medication *in vitro*.

Part 2 | DNA methylation algorithms and biological aging in schizophrenia

In **Chapter 5**, I describe how technical variation can affect DNAm-based predictors, in particular the performance of DNAm clocks. By implementing >100 data processing pipelines of commonly used DNAm methods, I provide a comprehensive evaluation of method performance and put forth guidelines and recommendations that minimize technical variation and maximize statistical power in analysis of DNAm-based biomarkers.

In **Chapter 6,** I describe a large meta-analysis of DNAm aging in schizophrenia. By analyzing four European cohorts and three different DNAm clocks, I describe how age and sex specific effects are primary drivers of differential DNAm aging in schizophrenia. Integration of schizophrenia polygenic risk identified that women in later adulthood who carry high genetic risk demonstrate the strongest age acceleration. This work suggests that specific and identifiable patient groups are at increased mortality risk as measured by the Levine DNAm clock. I describe how the biological specificity of the clock translates to epidemiological characteristics of the illness and how these may help in disease management and prevention in the clinic.

Part 3 | Discussion of results and conclusions

In **Chapter 7**, I discuss the research findings described in this dissertation, along with clinical implications for schizophrenia and new perspectives on our understanding of the illness. I place the work in a broader context through the lens of psychiatric and complex human trait genetics and describe limitations of my work alongside future research ideas.

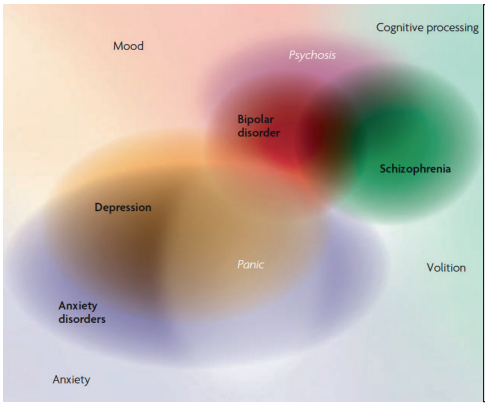
CHAPTER 2

The clinical presentation of schizophrenia and current genomic research standings

Schizophrenia is a severe and complex psychiatric disorder characterized by a heterogeneous combination of emotional, behavioral, and cognitive symptoms. The illness is caused by genetic or environmental factors, or both, and has an enormous burden on individuals affected, their families, and society (Kahn et al. 2015; Owen, Sawa, and Mortensen 2016). Compared with the general population, individuals with schizophrenia have a two- to threefold increased mortality risk (McGrath et al. 2008) and a life expectancy that is reduced by 15 years on average (Hjorthøj et al. 2017). The median lifetime morbid risk is estimated at ~0.7%, meaning that about 7 individuals per 1,000 will be affected, with substantial variation across geographical regions (McGrath et al. 2008). A large number of patients present long-term psychiatric symptoms and disability with the underlying biological mechanisms largely unknown and a cure yet to be found. Current treatment consists of prescription of antipsychotic medication combined with psychological and cognitive therapy and social support, while a need for more effective treatments remains. Among the foremost challenges is therefore to gain a deeper understanding of the causes and pathogenesis of the disorder in order to develop novel and effective treatments that are actionable in the clinic. Recent advances in human genetics and psychiatric research are starting to accelerate this process providing new avenues to explore.

Clinical presentation

The core features of schizophrenia are characterized by distorted thinking and perception and diminished emotional expression. Features are generally divided into "positive", "negative", and "cognitive" symptom domains. Positive symptoms are behaviors and thoughts that are usually not present, such as during a psychosis. An individual may experience a disconnect from reality presented through delusions, hallucinations, and disorganized behavior. Negative symptoms are characterized by a decrease in function in certain behavioral domains, such as social withdrawal, impaired motivation, affective flattening, and a reduction in spontaneous speech. Cognitive symptoms represent impairment over a wide range of cognitive functions, including but not limited to impaired attention, working memory dysfunction, and disrupted executive functions. The course of symptom expression can be either continuous or episodic with progressive or stable deficit. Patients can present one or more episodes with complete or incomplete remission. Clinical diagnosis of schizophrenia is generally assessed by criteria of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association 2013) or by criteria of the World Health Organisation's International Statistical Classification of Diseases and Related Health Problems (ICD-10) (World Health Organization, WHO Staff, and WHO 1992). Figure 1 shows an overview of symptom domains and diagnostic criteria for schizophrenia.



DSM-5:

A) Two or more of the following symptoms for up to 1 month unless treated successfully: delusions, hallucinations, disorganized speech, disorganized behavior, and negative symptoms (e.g. emotional flattening and avolition)
B) Level of functioning is significantly decreased in work, personal relationships and/or personal care.
C) Symptoms of the disorder persist at least 6 months.
D) Exclusion of schizo-affective disorder, unipolar and bipolar affective disorder. Symptoms cannot be attributed to the use of drugs or medication, or to a somatic disorder.
E) In the case of a pre-existing autism spectrum disorder, at least 1 month with prominent hallucinations or delusions.

ICD-10:

A) At least one of the following: echoing/insertion/withdrawal/broadcasting of thoughts, delusional perceptions, hallucinatory voices, impossible delusions of some kind.
B) Alternative to (A): at least two of the following: persistent hallucinations in any modality, incoherence or irrelevant speech, catatonic behavior, negative symptoms.
C) Present for most of the time for at least 1 month
D) Disorder is not caused by substance use or organic brain disease

Figure 1. A schematic overview of the definition of schizophrenia and how it exists around the classification of other major psychiatric illnesses. *The left panel is a color map with a schematic overview of phenotypic domains in which the symptomatology of the illness is visualized in relation to other major psychiatric illnesses and broader domains of emotions. The boundaries of these domains are illustrated as fluid representing the heterogeneity in symptoms and their partly overlapping characteristics. The image is adopted from Burmeister et al. 2008 (Burmeister, McInnis, and Zöllner 2008). The overview in the right panel is adopted from Kahn RS, et al., 2015 (Kahn et al. 2015) and lists the diagnostic criteria for schizophrenia as described in the DSM-V and ICD-10 guidelines.*

The first psychotic symptoms usually present themselves in late adolescence or early adulthood and often institute first contact with mental health services. A prodromal phase characterized by a decline in cognitive and social functioning can precede the first psychotic episode by many years (Kahn and Keefe 2013). Similar to its symptom heterogeneity, age at reported onset of the disorder has substantial variation with men reporting an earlier peak age at onset (early twenties) than women (mid-twenties) (Sham, MacLean, and Kendler 1994; Eranti et al. 2013). Individuals with an earlier age at onset are more likely to have more severe cognitive deficits and poorer global outcome than those with later onset (Rajji, Ismail, and Mulsant 2009). Those with later onset, especially female patients, have greater symptom overlap with mood disorders, such as major depressive and bipolar disorder (Sham, MacLean, and Kendler 1994; Hare et al. 2010). Overall, there is an inverse relationship between age at onset and familial risk of schizophrenia indicating that individuals with an earlier onset of symptoms are more likely to carry greater genetic risk for the illness (Sham, MacLean, and Kendler 1994; Hare et al. 2010).

The consequences of schizophrenia and burden of the illness

Although schizophrenia has a relatively low prevalence, it is ranked as one the most disabling illnesses globally (GBD 2016 Disease and Injury Incidence and Prevalence Collaborators

2017; Salomon et al. 2015). Not only do patients live many years with the consequences of the illness (Robinson et al. 2004; Harrison et al. 2001), they also report significant years of life lost due to increased risk of early death compared to the general population (Charlson et al. 2018; McGrath et al. 2008; Hjorthøj et al. 2017). Over the past decades, the burden of the illness has further increased. In part, this can be attributed to growth and ageing of the population while barriers to health care, such as access to medical care and social stigma, are likely contributors as well (Saha, Chant, and McGrath 2007). People diagnosed with schizophrenia are more likely to be unemployed, homeless, living in poverty, and overall report an excess prevalence of chronic physical illness at younger ages (Charlson et al. 2018; Strassnig et al. 2014). The comorbidity of schizophrenia with subsequent age-related somatic conditions, such as cardiovascular and respiratory illnesses and diabetes, significantly contribute to the excess early mortality (Laursen, Nordentoft, and Mortensen 2014; Olfson et al. 2015). In addition to the suffering of patients and their loved ones, schizophrenia also places a significant burden on the economy and society through direct cost of healthcare and indirect costs of cessation or reduction in work productivity (H. Y. Chong et al. 2016; Knapp, Mangalore, and Simon 2004). In the Netherlands, it is estimated that 2% of the national healthcare budget is spent on the treatment of schizophrenia, despite a prevalence of 0.6% (Evers, S M A, and A J H 1995). In the United States, the cost of schizophrenia is estimated at 2.5% of the total healthcare budget (Moscarelli, Rupp, and Sartorius 1996). These cost estimates are conservative as individuals diagnosed with schizophrenia are more likely to become homeless, unemployed, lose their access to the health care system, and die of subsequent comorbid conditions, thereby challenging accurate modeling of the burden of schizophrenia (Evers, S M A, and A J H 1995). Improving our understanding of the illness, both through the lens of its etiology as well as how it impacts the lives of individuals and the consequences of the illness they live with, is therefore needed to alleviate suffering, and improve the quality of life of those affected.

Schizophrenia genetics and the success of large-scale studies

Early genetic epidemiological studies using family and twin data have shown that genetic factors contribute to schizophrenia (Gottesman 1990; Sham, MacLean, and Kendler 1994). Studies of twin meta-analysis and population twin registries report an overall heritability of up to 81% (Polderman et al. 2015; Hilker et al. 2018; Sullivan, Kendler, and Neale 2003), indicating a substantial but not exclusive contribution of genetic factors. While not the only risk factor, having a first-degree relative with schizophrenia is one of the most important ones (Murray et al. 2002). Monozygotic twins for example report diagnostic concordance rates of about 33%, while for dizygotic twins it is reported at 7% (Hilker et al. 2018; Kendler and Robinette 1983), further indicating that illness vulnerability is not solely determined by genetic factors. There is also consistent evidence of shared environmental influences on the risk for schizophrenia, which is estimated at 11% (Sullivan, Kendler, and Neale 2003). Figure 2 shows an overview of the history of schizophrenia genetic research over time.

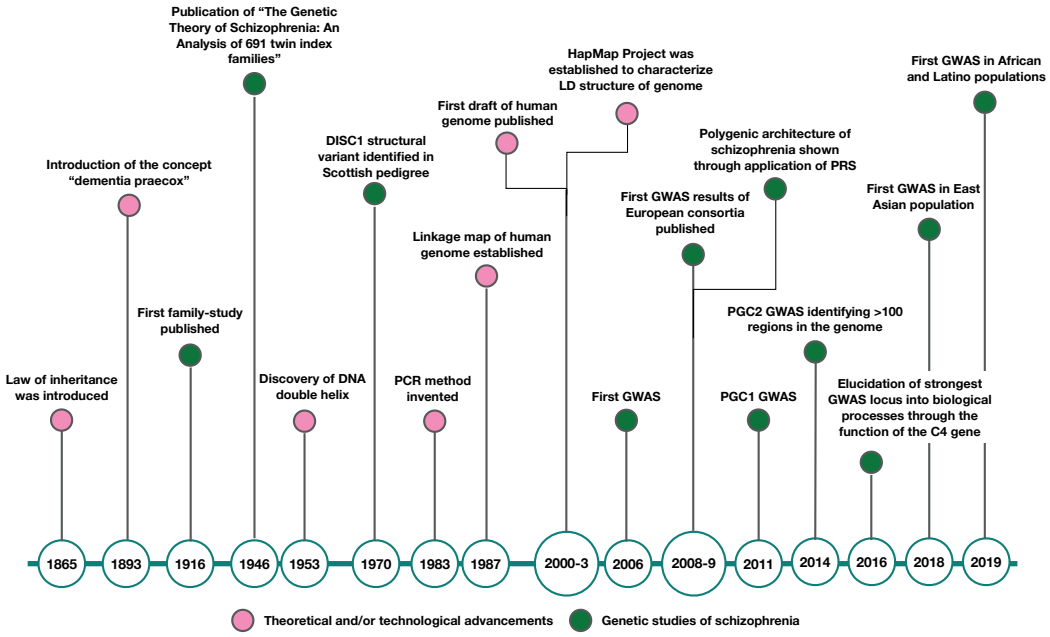


Figure 2. A timeline of the history of schizophrenia genetic research. *The recent acceleration of our understanding of the genetic architecture of schizophrenia is the result of decades of research advancements, both in the theory of phenotype definition and in technological innovations that allowed large-scale genetic studies to be conducted. While early family studies of schizophrenia indicated heredity to play a role, it was not until our understanding of the genome matured and research groups joint efforts in international consortia that enabled characterization of the genetic architecture of schizophrenia. This is embodied by the work of the Psychiatric Genomics Consortium (PGC). First through their GWAS efforts in European populations, which are now extended to Asian, African, and Latin-American populations as well.*

Given its high heritability and relatively more distinct clinical presentation compared to other psychiatric illnesses, schizophrenia research has been a leading undertaking in psychiatric genetics. This is embodied by the efforts of the Psychiatric Genomics Consortium (PGC), a global consortium representing hundreds of scientists from across the world. The PGC conducted a landmark genetic study (a GWAS) of schizophrenia comprising 36,989 cases and 113,075 controls and identified 108 independent regions across the genome as significantly associated with the illness (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). This work was published around the start of my research on schizophrenia biology. A later GWAS with increased sample size reports 145 independent loci (Pardiñas et al. 2018) and the most recent iteration even 270 loci (Consortium et al., n.d.). As GWAS examines only common genetic variation, studies of lower frequency variation have been conducted as well to advance our understanding of the genetic causes of the illness. Copy number variation (CNV), which are deletions or duplications of

stretches of DNA, are increased in schizophrenia compared to the general population (Marshall et al. 2017). Significantly associated CNVs collectively explain 0.85% of the variance in schizophrenia liability in the PGC sample. In comparison, 3.4% of the variance in disease liability is explained for the 108 significant GWAS loci. In addition to CNVs, rare single variants that are predicted to be damaging or deleterious in their outcome are detected at increased frequency in schizophrenia as well (Purcell et al. 2014; Olde Loohuis et al. 2015; Singh et al. 2017). Figure 3 shows an overview of schizophrenia genetic risk variants across the allele frequency sprectrum. Together, these results unequivocally demonstrate the success of using large-scale genetic studies to identify genetic risk of schizophrenia. Several important lessons have emerged from these findings so far.

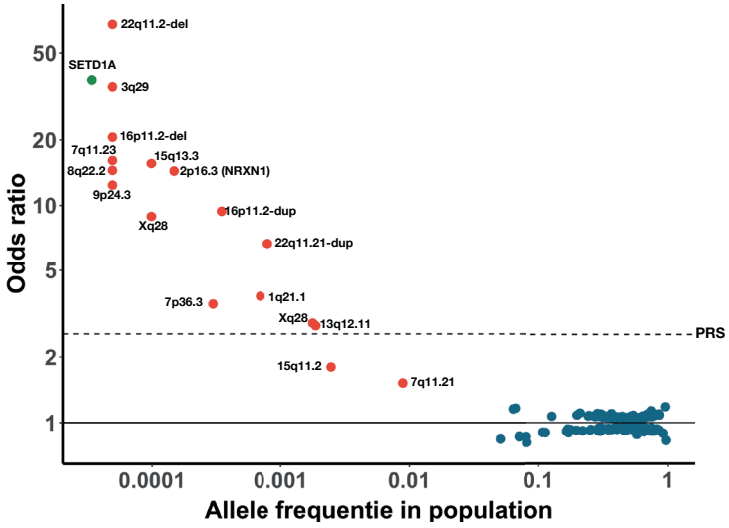


Figure 3. An overview of schizophrenia genetic risk variants across the allele frequency spectrum. *For each genetic variant or locus, the strength of association (odds ratio) with schizophrenia is presented alongside the frequency of the risk allele in the control population. Only genome-wide significant variants are shown. Blue data points represent 176 common variant associations from a meta-analysis of a large European GWAS of schizophrenia (Ripke et al., 2014). Red data points show the association of 17 copy number variants reported in Marshall et al., 2017. The green data point shows the association of a rare loss-of-function variant based on exome sequencing data (Singh et al., 2016). Both axes are transformed to a logarithmic scale. For comparison, the dotted horizontal line shows the association of aggregated polygenic risk computed from common variant associations (based on Ripke et al., 2014). The odds ratio is derived from Ori et al., 2019 (BiorxivID: 727859).*

Lesson 1: polygenicity

One of the main insights from GWAS has been the confirmation that the genetic architecture of schizophrenia is polygenic, meaning that many distinct regions in the genome collectively contribute to an individual's genetic propensity to develop the illness at some point in their life. For example, while no single genetic variant from the GWAS explains more than 0.1% of

schizophrenia risk, ~10,000 of SNPs together explain up to 18.4% of the genetic risk (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Martin et al. 2019). In line with Fisher's infinitesimal model, which proposes that a large number of independent genetic loci contribute additively to continuous phenotypic variation, each variant thus explains a small fraction of the overall heritability. This has important implications for functional investigations of schizophrenia biology. Results from candidate gene studies, especially those using full gene knockouts, for example, need to be carefully evaluated before findings are extrapolated to gain mechanistic insights into schizophrenia (Farrell et al. 2015). Future experimental follow-up studies furthermore need to pursue new strategies that allow for modeling many small gene effects collectively in combination with state-of-the-art laboratory techniques that measure molecular and cellular readouts.

Lesson 2: pleiotropy

A second key insight is that there is significant widespread shared heritability between major psychiatric disorders (Lee et al. 2013; O'Donovan and Owen 2016). In other words, psychiatric genetic risk is pleiotropic and genetic variants that confer risk for schizophrenia also confer risk for bipolar disorder and major depressive disorder, for example. Based on common variant analysis, schizophrenia shares a high genetic overlap with bipolar disorder (~68%) and a moderate overlap with major depressive disorder (~43%) (Lee et al. 2013; Brainstorm Consortium et al. 2018). Pleiotropy has also been reported for rare variants. CNVs that confer risk for schizophrenia also affect a range of childhood neurodevelopmental disorders, such as autism spectrum disorder and intellectual disability (Malhotra and Sebat 2012). A finding that is also supported by rare coding variants (O'Donovan and Owen 2016). It has been theorized that neurodevelopmental disorders, including schizophrenia, lay on an etiological and neurodevelopmental continuum, where genetic variants are shared but the phenotypic expressivity across disorders is variable (Owen and O'Donovan 2017). While overlapping mechanisms are likely at work, how pleiotropy and variable expressivity operate on a biological level remains an important open question for future research.

Lesson 3: biological insights

A third key insight stems from genes and pathways that have been identified through genetic variants associated with schizophrenia. Early studies of rare disruptive CNVs in schizophrenia reported enrichment in neurodevelopmental and synaptic gene sets (Walsh et al. 2008; Glessner et al. 2010; Kirov et al. 2012), findings that were later replicated by efforts with larger sample sizes (Pocklington et al. 2015; Szatkiewicz et al. 2014; Marshall et al. 2017). Neurodevelopmental pathways are also highlighted by genome-wide burden analysis of rare single deleterious variants, which are enriched in schizophrenia (Olde Loohuis et al. 2015). Genes involved in synaptic function are furthermore implicated through analyses of common variants, in addition to pathways of neuronal functioning and the immune system (Aberg et al. 2013; Stefansson et al. 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). More sophisticated analyses that better leverage genome-wide information found schizophrenia heritability to be significantly enriched for central nervous system tissues

with the strongest signal observed in fetal brain chromatin annotations (Finucane et al. 2015), again highlighting early brain development. Together these findings not only demonstrate that schizophrenia genetic risk concentrates in specific biological annotations, they also show that results of rare and common variant analyses are starting to converge to similar pathways and thereby strengthening the evidence for their role in schizophrenia biology. To pave the way for functional follow-up studies of GWAS associations, experimental model systems that are relevant to the biology of the disorder are needed. Lack of access to the brain and our inability to measure neurobiological markers *in vivo* is currently limiting fundamental and translational brain-related research. Alternative approaches that capture dimensions of human brain function can help accelerate our understanding of the etiology of schizophrenia.

Lesson 4: clinical implementation and utility so far

In addition to improving our understanding of the disease biology, the identification of a large number of independent genetic variants associated with schizophrenia initiated new efforts to utilize genetic information to advance actionability in the clinic. Two examples of research for this purpose are drug repositioning and disease risk predictions. The process of finding and developing new clinical uses for existing licensed drugs outside their initial medical domain is known as repositioning (Ashburn and Thor 2004; C. R. Chong and Sullivan 2007). While antipsychotics can provide effective treatment for schizophrenia, they do not alleviate all symptoms and often contribute to serious side-effects (Leucht et al. 2013), leading to reduced adherence and lower efficacy. Why antipsychotics show better results for some patients but not all remains unclear. A better understanding of how drug and disease mechanisms interact is therefore needed. New insights may help improve the efficacy of existing drugs and help design new drugs. Genomic data can be used for this purpose. GWAS results can for example prioritize drugs and their biological targets for a given phenotype and help guide drug discovery (Héléna A. Gaspar, Hübel, and Breen 2019). For schizophrenia, the dopamine D2 receptor gene (DRD2), a main target of antipsychotics, lies within a genome-wide significant locus identified by GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). Target gene-sets of dopamine receptor antagonists are in addition found to be enriched for schizophrenia GWAS signal (de Jong et al. 2016). In general, schizophrenia genetic findings so far point to overlap with some targets from known antipsychotics but also new drugs like antiepileptics and calcium channel blockers, offering possible leads for developing new therapeutics (H. A. Gaspar and Breen 2017; So et al. 2017; Ruderfer et al. 2016). Such an overlap between drug target gene sets and genes that harbor elevated genetic risk for the disease would not be unique to schizophrenia. For coronary artery disease for example, several primary drug target genes, such as the downstream molecular targets of statins, have now also been identified by GWAS (Turner et al. 2018). Similar observations have been reported for diabetes mellitus and rheumatoid arthritis. More in general, GWAS genes tend to be closer to drug target genes in gene network analysis of biological pathways (Cao and Moult 2014). This suggests that the study of molecular targets of medication used to treat the illness can shed light on both the biology of adverse effects of the drug as well as possible causal genes and pathways involved in the disease. As our knowledge of the biology of antipsychotics is still limited, integrating genes and pathways associated with the drug with genetic risk of schizophrenia may yield new insights.

A second line of research aiming to advance actionability in the clinic using information from GWAS has been through investigations of polygenic risk score (PRS) predictions. PRSs, also called polygene scores or genetic values, are a weighted sum of risk alleles of SNPs and represent a single composite value that quantifies the cumulative genetic load of common variant associations for an individual (Torkamani, Wineinger, and Topol 2018; Janssens 2019). Early disease detection, subsequent prevention and medical interventions are integral components of advancing human health. The potential of estimating a probabilistic susceptibility of an individual to disease is therefore driving a rapidly progressing field of research investigating the personal and clinical utility of PRS. Particularly in psychiatry, this has created hope for a new classification system based on biological validity (Kapur, Phillips, and Insel 2012). Schizophrenia was one of the first human diseases for which PRS, calculated from effect sizes of the GWAS, was shown to be able to discriminate between individuals diagnosed with the illness and those without the illness (International Schizophrenia Consortium et al. 2009). While this finding has now been robustly replicated and shown for other psychiatric illnesses as well (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Martin et al. 2019), their low predictive value and the complexity of polygenic inheritance, among other challenges, currently limit the immediate utility of these results (Torkamani, Wineinger, and Topol 2018; Janssens 2019). There is nevertheless significant promise for future clinical utility as sample sizes and ancestral diversity of genetic studies of schizophrenia continue to increase (Lam et al. 2019; Sullivan et al. 2017; Gulsuner et al. 2020) and PRS methodology is further refined (Martin et al. 2019). In parallel, investigations of other genomic biomarkers, such as gene expression and DNA methylation measures, may provide complementary value to further stratify patients into meaningful subgroups across conventional diagnostic boundaries.

DNA methylation as a tally for health and disease

Schizophrenia is a complex illness with a multifactorial etiology and variable disease trajectory. Many factors, such as DNA sequence, epigenetic DNA modifications, gene expression and protein differences, changes in cellular profiles, environmental factors, stochastic factors, and the complex and dynamic interaction between these are thought to act in concert to influence the outcome of the illness (Haque, Gottesman, and Wong 2009; Flint and Munafò 2014; Kahn et al. 2015). Measuring these factors in the context of the illness is an important step towards understanding their contribution. DNA methylation (DNAm), the covalent attachment of a methyl (CH_3) to the DNA, is a measurable genomic signature and a form of epigenetic modification and regulation. In mammals, DNAm usually occurs at 5-methylcytosine (5mC) at cytosine-guanine dinucleotides (CpGs) sites in the genome. DNAm has been described in diverse roles in human development and disease, including transcription activation, X-chromosome inactivation, genomic imprinting, patterning waves during development, and many other key regulatory and cellular processes, including aging (Greenberg and Bourc'his 2019; Horvath and Raj 2018). Recent technological advancements allow for genome-wide measurement of DNAm in a high-throughput manner thereby enabling large-scale studies to be conducted (Beck 2010).

Over the past years, several observations have been established that suggest that DNAm is a meaningful biological signature to study in schizophrenia. First, DNAm can be influenced

by both internal and external cues. Twin studies have reported the contribution of additive genetic effects, shared environmental effects, and unshared (or unique) environment to DNAm variation at 16%, 17%, and 67% respectively (Hannon et al. 2018). This architecture fits well with the multifactorial origin of complex phenotypes, like schizophrenia. Secondly, DNAm measured in blood tissue tracks health and lifestyle exposure that are associated with schizophrenia, such as smoking, body-mass-index (BMI), and diet (Gao et al. 2015; Joehanes et al. 2016; Zhang and Kutateladze 2018; Wahl et al. 2017). Large epigenome-wide association studies (EWAS) have found wide-spread changes across the genome as a consequence of such exposures. These associated DNAm changes furthermore predict future development of health and disease offering opportunities for clinical utility (Wahl et al. 2017; Sugden et al. 2019). The third observation comes from EWA studies of schizophrenia. In a large cohort, >350 CpG sites were found to be associated with the illness. Overall the findings could be differentiated between DNAm changes that were associated with schizophrenia through smoking and those that were independent of smoking behavior and other indirect effects (Hannon et al. 2016). Indeed, measured DNAm signatures are a composition of intrinsic and extrinsic factors that have impacted an individual's biological state thus far (Teschendorff and Relton 2018; Lappalainen and Greally 2017). While reverse-causation and interpretation of results remain main challenges of the outcome of DNAm studies, the landscape of DNAm does offer new opportunities to study health and disease. This is embodied by the fourth and final observation; the development of DNA methylation clocks as a biomarker of aging and a tally of health and disease, including mortality risk.

The phenomenon of aging can be described as a cumulative result of biological processes over time. Chronological age is one of the strongest, if not the strongest, risk factor for functional impairments, disease development, and mortality. There is however significant heterogeneity in the aging process of individuals in the populations (Lowsky et al. 2014). Tracking the rate of aging above and beyond chronological age, which is an imperfect surrogate measure of the aging process, has therefore been of major interest for a long time (Baker and Sprott 1988). In theory, a marker sensitive to the pace of biological aging can be an important clinical tool for disease prevention and management (see Figure 4). Several molecular markers have been extensively studied, such as levels of P16 and telomere length, both markers of cellular senescence (Waaijer et al. 2012; Epel et al. 2008). Singular predictors however capture specific processes of aging and are less predictive of global aging and all-cause mortality risk than composite biomarkers (Jylhävä, Pedersen, and Hägg 2017). Recent approaches using -omics data, suggest that genome-wide profiles of gene expression and epigenetic marks can be aggregated to a single value of aging that better captures the complexities of the aging process (Peters et al. 2015; Horvath and Raj 2018). In particular, DNA methylation-based biomarkers of aging, also called "epigenetic clocks" or "DNA methylation clocks" (DNAm clocks), show great promise in tracking the pace of biological aging (Horvath 2013; Hannum et al. 2013; Levine et al. 2018). DNAm clocks outperform traditional biomarkers of aging (Jylhävä, Pedersen, and Hägg 2017), are significant predictors of age-related conditions and mortality (Chen et al. 2016; Levine et al. 2018), and have been associated with several diseases (Horvath et al. 2014), including psychiatric conditions (Boks et al. 2015; Wolf et al. 2019; Han et al. 2018). These clocks may thus offer an opportunity to study how biological aging is impacted in schizophrenia.

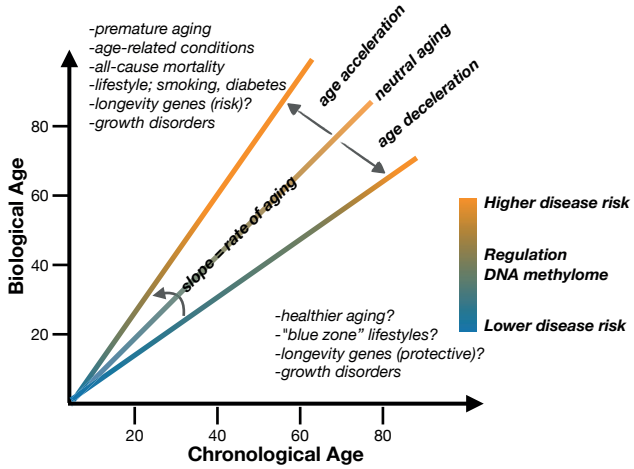


Figure 4. The pace of biological aging is defined by both intrinsic and extrinsic factors. *The relationship between biological age (y-axis) and chronological age (x-axis) is visually shown. The pace of aging can be measured by the slope of the relationship. When biological age tracks with chronological age the pace of aging is "neutral". However, when biological age is higher than chronological age, biological age is accelerated. Similarly, when biological age is lower than chronological age, biological age is decelerated. Various factors that contribute or that are related to biological aging are listed in the figure as well.*

There are long-standing epidemiological observations that describe associations of schizophrenia with age-related disabilities and morbidities at younger ages. Dating back to the work of Emil Kraepelin and others at the beginning of the 1900s, premature progressive deterioration of cognitive functions (dementia praecox) during early adulthood is reported as a core feature of what is now known as schizophrenia (Kraepelin 1921). In addition to cognitive decline, excess obesity-related and metabolic abnormalities, decrease in cardiovascular fitness, impaired motor functions, among other age-related conditions (Strassnig et al. 2014). The most striking evidence that supports a theory of accelerated aging, comes from studies of excess mortality in schizophrenia. That is, even at younger ages, individuals diagnosed with schizophrenia show higher rates of all-cause mortality compared to the general population (Laursen, Nordentoft, and Mortensen 2014; Olfson et al. 2015). Processes of biological aging may therefore be accelerated in schizophrenia, either through an increased prevalence of age-related conditions with effects that compound over time or as a more integrated part of the pathophysiology of the illness (Kirkpatrick et al. 2008). As epigenetic signatures can be modifiable (Sugden et al. 2019), DNAm-based predictors may have clinical utility. Quantification of biological aging can help with identification of at-risk individuals or even prevention of age-related diseases (Belsky et al. 2015; Field et al. 2018). As the burden of age-related diseases continues to rise, early detection and subsequent opportunities for interventions before disabilities and co-morbidities become established will be important (Moffitt and Caspi 2019; Taylor and Reynolds 2020). DNAm clocks are now emerging as promising tools for screening and intervention and offer new opportunities to study the phenomenon of aging, and possibly signatures of molecular consequences, in SCZ.

References

Aberg, Karolina A., Youfang Liu, Jozsef Bukszár, Joseph L. McClay, Amit N. Khachane, Ole A. Andreassen, Douglas Blackwood, et al. 2013. "A Comprehensive Family-Based Replication Study of Schizophrenia Genes." JAMA Psychiatry 70 (6): 573–81.

American Psychiatric Association. 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Ashburn, Ted T., and Karl B. Thor. 2004. "Drug Repositioning: Identifying and Developing New Uses for Existing Drugs." Nature Reviews. Drug Discovery 3 (8): 673–83.

Baker, G. T., 3rd, and R. L. Sprott. 1988. "Biomarkers of Aging." Experimental Gerontology 23 (4-5): 223–39.

Beck, Stephan. 2010. "Taking the Measure of the Methylome." Nature Biotechnology.

Belsky, Daniel W., Avshalom Caspi, Renate Houts, Harvey J. Cohen, David L. Corcoran, Andrea Danese, Honalee Harrington, et al. 2015. "Quantification of Biological Aging in Young Adults." Proceedings of the National Academy of Sciences of the United States of America 112 (30): E4104–10.

Boks, Marco P., Hans C. van Mierlo, Bart P. F. Rutten, Timothy R. D. J. Radstake, Lot De Witte, Elbert Geuze, Steve Horvath, et al. 2015. "Longitudinal Changes of Telomere Length and Epigenetic Age Related to Traumatic Stress and Post-Traumatic Stress Disorder." Psychoneuroendocrinology 51 (January): 506–12.

Brainstorm Consortium, Verneri Anttila, Brendan Bulik-Sullivan, Hilary K. Finucane, Raymond K. Walters, Jose Bras, Laramie Duncan, et al. 2018. "Analysis of Shared Heritability in Common Disorders of the Brain." Science 360 (6395).

Burmeister, Margit, Melvin G. McInnis, and Sebastian Zöllner. 2008. "Psychiatric Genetics: Progress amid Controversy." Nature Reviews. Genetics 9 (7): 527–40.

Cao, Chen, and John Moult. 2014. "GWAS and Drug Targets." BMC Genomics 15 Suppl 4 (May): S5.

Charlson, Fiona J., Alize J. Ferrari, Damian F. Santomauro, Sandra Diminic, Emily Stockings, James G. Scott, John J. McGrath, and Harvey A. Whiteford. 2018. "Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016." Schizophrenia Bulletin 44 (6): 1195–1203.

Chen, Brian H., Riccardo E. Marioni, Elena Colicino, Marjolein J. Peters, Cavin K. Ward-Caviness, Pei-Chien Tsai, Nicholas S. Roetker, et al. 2016. "DNA Methylation-Based Measures of Biological Age: Meta-Analysis Predicting Time to Death." Aging 8 (9): 1844–65.

Chong, Curtis R., and David J. Sullivan. 2007. "New Uses for Old Drugs." Nature 448 (7154): 645–46.

Chong, Huey Yi, Siew Li Teoh, David Bin-Chia Wu, Surachai Kotirum, Chiun-Fang Chiou, and Nathorn Chaiyakunapruk. 2016. "Global Economic Burden of Schizophrenia: A Systematic Review." Neuropsychiatric Disease and Treatment 12 (February): 357–73.

Consortium, Schizophrenia Working Group of The Psychiatric Genomics, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Stephan Ripke, James T. R. Walters, and Michael C. O'Donovan. 2020. "Mapping Genomic Loci Prioritises Genes and Implicates Synaptic Biology in Schizophrenia." https://doi.org/10.1101/2020.09.12.20192922.

Epel, Elissa S., Sharon Stein Merkin, Richard Cawthon, Elizabeth H. Blackburn, Nancy E. Adler, Mark J. Pletcher, and Teresa E. Seeman. 2008. "The Rate of Leukocyte Telomere Shortening Predicts Mortality from Cardiovascular Disease in Elderly Men: A Novel Demonstration." Aging.

Eranti, S. V., J. H. MacCabe, H. Bundy, and R. M. Murray. 2013. "Gender Difference in Age at Onset of Schizophrenia: A Meta-Analysis." Psychological Medicine 43 (1): 155–67.

Evers, S. M. A. A., S M A, and A J H. 1995. "Costs of Schizophrenia in the Netherlands." Schizophrenia Bulletin.

Farrell, M. S., T. Werge, P. Sklar, M. J. Owen, R. A. Ophoff, M. C. O'Donovan, A. Corvin, S. Cichon, and P. F. Sullivan. 2015. "Evaluating Historical Candidate Genes for Schizophrenia." Molecular Psychiatry 20 (5): 555–62.

Field, Adam E., Neil A. Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D. Adams. 2018. "DNA Methylation Clocks in Aging: Categories, Causes, and Consequences." Molecular Cell 71 (6): 882–95.

Finucane, Hilary K., ReproGen Consortium, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, et al. 2015. "Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics." Nature Genetics.

Flint, Jonathan, and Marcus Munafò. 2014. "Genesis of a Complex Disease." Nature.

Gao, Xu, Min Jia, Yan Zhang, Lutz Philipp Breitling, and Hermann Brenner. 2015. "DNA Methylation Changes of Whole Blood Cells in Response to Active Smoking Exposure in Adults: A Systematic Review of DNA Methylation Studies." Clinical Epigenetics 7 (October): 113.

Gaspar, H. A., and G. Breen. 2017. "Drug Enrichment and Discovery from Schizophrenia Genome-Wide Association Results: An Analysis and Visualisation Approach." Scientific Reports 7 (1): 12460.

Gaspar, Héléna A., Christopher Hübel, and Gerome Breen. 2019. "Drug Targetor: A Web Interface to Investigate the Human Druggome for over 500 Phenotypes." Bioinformatics 35 (14): 2515–17.

GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. 2017. "Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 328 Diseases and Injuries for 195 Countries, 1990-2016: A Systematic Analysis for the Global Burden of Disease Study 2016." The Lancet 390 (10100): 1211–59.

Glessner, Joseph T., Muredach P. Reilly, Cecilia E. Kim, Nagahide Takahashi, Anthony Albano, Cuiping Hou, Jonathan P. Bradfield, et al. 2010. "Strong Synaptic Transmission Impact by Copy Number Variations in Schizophrenia." Proceedings of the National Academy of Sciences of the United States of America 107 (23): 10584–89.

Gottesman, Irving I. 1990. Schizophrenia Genesis: The Origins of Madness. W. H. Freeman. Greenberg, Maxim V. C., and Deborah Bourc'his. 2019. "The Diverse Roles of DNA Methylation in Mammalian Development and Disease." Nature Reviews. Molecular Cell Biology 20 (10): 590–607.

Gulsuner, S., D. J. Stein, E. S. Susser, G. Sibeko, A. Pretorius, T. Walsh, L. Majara, et al. 2020. "Genetics of Schizophrenia in the South African Xhosa." Science 367 (6477): 569–73.

Han, Laura K. M., Moji Aghajani, Shaunna L. Clark, Robin F. Chan, Mohammad W. Hattab, Andrey A. Shabalin, Min Zhao, et al. 2018. "Epigenetic Aging in Major Depressive Disorder." The American Journal of Psychiatry 175 (8): 774–82.

Hannon, Eilis, Emma Dempster, Joana Viana, Joe Burrage, Adam R. Smith, Ruby Macdonald, David St Clair, et al. 2016. "An Integrated Genetic-Epigenetic Analysis of Schizophrenia: Evidence for Co-Localization of Genetic Associations and Differential DNA Methylation." Genome Biology 17 (1): 176.

Hannon, Eilis, Olivia Knox, Karen Sugden, Joe Burrage, Chloe C. Y. Wong, Daniel W. Belsky, David L. Corcoran, et al. 2018. "Characterizing Genetic and Environmental Influences on Variable DNA Methylation Using Monozygotic and Dizygotic Twins." PLoS Genetics 14 (8): e1007544.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." Molecular Cell 49 (2): 359–67.

Haque, F. Nipa, Irving I. Gottesman, and Albert H. C. Wong. 2009. "Not Really Identical: Epigenetic Differences in Monozygotic Twins and Implications for Twin Studies in Psychiatry." American Journal of Medical Genetics. Part C, Seminars in Medical Genetics 151C (2): 136– 41.

Hare, Elizabeth, David C. Glahn, Albana Dassori, Henriette Raventos, Humberto Nicolini, Alfonso Ontiveros, Rolando Medina, et al. 2010. "Heritability of Age of Onset of Psychosis in Schizophrenia." American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 153B (1): 298–302.

Harrison, G., K. Hopper, T. Craig, E. Laska, C. Siegel, J. Wanderling, K. C. Dube, et al. 2001. "Recovery from Psychotic Illness: A 15- and 25-Year International Follow-up Study." The British Journal of Psychiatry: The Journal of Mental Science 178 (June): 506–17.

Hilker, Rikke, Dorte Helenius, Birgitte Fagerlund, Axel Skytthe, Kaare Christensen, Thomas M. Werge, Merete Nordentoft, and Birte Glenthøj. 2018. "Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register." Biological Psychiatry 83 (6): 492–98.

Hjorthøj, Carsten, Anne Emilie Stürup, John J. McGrath, and Merete Nordentoft. 2017. "Years of Potential Life Lost and Life Expectancy in Schizophrenia: A Systematic Review and Meta-Analysis." The Lancet. Psychiatry 4 (4): 295–301.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." Genome Biology 14 (10): R115.
Horvath, Steve, Wiebke Erhart, Mario Brosch, Ole Ammerpohl, Witigo von Schönfels, Markus Ahrens, Nils Heits, et al. 2014. "Obesity Accelerates Epigenetic Aging of Human Liver." Proceedings of the National Academy of Sciences, 201412759.

Horvath, Steve, and Kenneth Raj. 2018. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing." Nature Reviews. Genetics 19 (6): 371–84.

International Schizophrenia Consortium, Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, and Pamela Sklar. 2009. "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder." Nature 460 (7256): 748–52.

Janssens, A. Cecile J. W. 2019. "Validity of Polygenic Risk Scores: Are We Measuring What We Think We Are?" Human Molecular Genetics 28 (R2): R143–50.

Joehanes, Roby, Allan C. Just, Riccardo E. Marioni, Luke C. Pilling, Lindsay M. Reynolds, Pooja R. Mandaviya, Weihua Guan, et al. 2016. "Epigenetic Signatures of Cigarette Smoking." Circulation. Cardiovascular Genetics 9 (5): 436–47.

Jong, Simone de, Lewis R. Vidler, Younes Mokrab, David A. Collier, and Gerome Breen. 2016. "Gene-Set Analysis Based on the Pharmacological Profiles of Drugs to Identify Repurposing Opportunities in Schizophrenia." Journal of Psychopharmacology 30 (8): 826–30.

Jylhävä, Juulia, Nancy L. Pedersen, and Sara Hägg. 2017. "Biological Age Predictors." EBioMedicine.

Kahn, René S., and Richard S. E. Keefe. 2013. "Schizophrenia Is a Cognitive Illness: Time for a Change in Focus." JAMA Psychiatry 70 (10): 1107–12.

Kahn, René S., Iris E. Sommer, Robin M. Murray, Andreas Meyer-Lindenberg, Daniel R. Weinberger, Tyrone D. Cannon, Michael O'Donovan, et al. 2015. "Schizophrenia." Nature Reviews. Disease Primers 1 (November): 15067.

Kapur, S., A. G. Phillips, and T. R. Insel. 2012. "Why Has It Taken so Long for Biological Psychiatry to Develop Clinical Tests and What to Do about It?" Molecular Psychiatry.

Kendler, K. S., and C. D. Robinette. 1983. "Schizophrenia in the National Academy of Sciences-National Research Council Twin Registry: A 16-Year Update." The American Journal of Psychiatry 140 (12): 1551–63.

Kirkpatrick, Brian, Erick Messias, Philip D. Harvey, Emilio Fernandez-Egea, and Christopher R. Bowie. 2008. "Is Schizophrenia a Syndrome of Accelerated Aging?" Schizophrenia Bulletin 34 (6): 1024–32.

Kirov, G., A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, et al. 2012. "De Novo CNV Analysis Implicates Specific Abnormalities of Postsynaptic Signalling Complexes in the Pathogenesis of Schizophrenia." Molecular Psychiatry 17 (2): 142–53.

Knapp, M., R. Mangalore, and J. Simon. 2004. "The Global Costs of Schizophrenia." Schizophrenia Bulletin.

Kraepelin, Emil. 1921. "DEMENTIA PRAECOX AND PARAPHRENIA." The Journal of Nervous and Mental Disease.

Lam, Max, Chia-Yen Chen, Zhiqiang Li, Alicia R. Martin, Julien Bryois, Xixian Ma, Helena Gaspar, et al. 2019. "Comparative Genetic Architectures of Schizophrenia in East Asian and European Populations." Nature Genetics 51 (12): 1670–78.

Lappalainen, Tuuli, and John M. Greally. 2017. "Associating Cellular Epigenetic Models with Human Phenotypes." Nature Reviews. Genetics 18 (7): 441–51.

Laursen, Thomas Munk, Merete Nordentoft, and Preben Bo Mortensen. 2014. "Excess Early Mortality in Schizophrenia." Annual Review of Clinical Psychology 10: 425–48.

Lee, S. Hong, Stephan Ripke, Benjamin M. Neale, Stephen V. Faraone, Shaun M. Purcell, Roy H. Perlis, Bryan J. Mowry, et al. 2013. "Genetic Relationship between Five Psychiatric Disorders Estimated from Genome-Wide SNPs." Nature Genetics 45 (9): 984–94.

Leucht, Stefan, Andrea Cipriani, Loukia Spineli, Dimitris Mavridis, Deniz Orey, Franziska Richter, Myrto Samara, et al. 2013. "Comparative Efficacy and Tolerability of 15 Antipsychotic Drugs in Schizophrenia: A Multiple-Treatments Meta-Analysis." The Lancet 382 (9896): 951– 62.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." Aging 10 (4): 573–91.

Lowsky, David J., S. Jay Olshansky, Jay Bhattacharya, and Dana P. Goldman. 2014. "Heterogeneity in Healthy Aging." The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences 69 (6): 640–49.

Malhotra, Dheeraj, and Jonathan Sebat. 2012. "CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics." Cell 148 (6): 1223–41.

Marshall, Christian R., Daniel P. Howrigan, Daniele Merico, Bhooma Thiruvahindrapuram, Wenting Wu, Douglas S. Greer, Danny Antaki, et al. 2017. "Contribution of Copy Number Variants to Schizophrenia from a Genome-Wide Study of 41,321 Subjects." Nature Genetics 49 (1): 27–35.

Martin, Alicia R., Mark J. Daly, Elise B. Robinson, Steven E. Hyman, and Benjamin M. Neale. 2019. "Predicting Polygenic Risk of Psychiatric Disorders." Biological Psychiatry 86 (2): 97– 109.

McGrath, John, Sukanta Saha, David Chant, and Joy Welham. 2008. "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality." Epidemiologic Reviews 30 (May): 67–76.

Moffitt, Terrie E., and Avshalom Caspi. 2019. "Psychiatry's Opportunity to Prevent the Rising Burden of Age-Related Disease." JAMA Psychiatry.

Moscarelli, M., A. Rupp, and Norman Sartorius. 1996. Handbook of Mental Health Economics and Health Policy, Schizophrenia. John Wiley & Son Limited.

Murray, Robin M., Peter B. Jones, Ezra Susser, Jim Van Os, and Mary Cannon. 2002. The Epidemiology of Schizophrenia. Cambridge University Press.

O'Donovan, Michael C., and Michael J. Owen. 2016. "The Implications of the Shared Genetics of Psychiatric Disorders." Nature Medicine 22 (11): 1214–19.

Olde Loohuis, Loes M. Olde, Jacob A. S. Vorstman, Anil P. Ori, Kim A. Staats, Tina Wang, Alexander L. Richards, Ganna Leonenko, et al. 2015. "Genome-Wide Burden of Deleterious Coding Variants Increased in Schizophrenia." Nature Communications 6: 7501.

Olfson, Mark, Tobias Gerhard, Cecilia Huang, Stephen Crystal, and T. Scott Stroup. 2015. "Premature Mortality Among Adults With Schizophrenia in the United States." JAMA Psychiatry 72 (12): 1172–81.

Owen, Michael J., and Michael C. O'Donovan. 2017. "Schizophrenia and the Neurodevelopmental Continuum:evidence from Genomics." World Psychiatry: Official Journal of the World Psychiatric Association 16 (3): 227–35.

Owen, Michael J., Akira Sawa, and Preben B. Mortensen. 2016. "Schizophrenia." The Lancet 388 (10039): 86–97.

Pardiñas, Antonio F., Peter Holmans, Andrew J. Pocklington, Valentina Escott-Price, Stephan Ripke, Noa Carrera, Sophie E. Legge, et al. 2018. "Common Schizophrenia Alleles Are Enriched in Mutation-Intolerant Genes and in Regions under Strong Background Selection." Nature Genetics 50 (3): 381–89.

Peters, Marjolein J., Roby Joehanes, Luke C. Pilling, Claudia Schurmann, Karen N. Conneely, Joseph Powell, Eva Reinmaa, et al. 2015. "The Transcriptional Landscape of Age in Human Peripheral Blood." Nature Communications 6 (October): 8570.

Pocklington, Andrew J., Elliott Rees, James T. R. Walters, Jun Han, David H. Kavanagh, Kimberly D. Chambert, Peter Holmans, et al. 2015. "Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia." Neuron 86 (5): 1203–14.

Polderman, Tinca J. C., Beben Benyamin, Christiaan A. de Leeuw, Patrick F. Sullivan, Arjen van Bochoven, Peter M. Visscher, and Danielle Posthuma. 2015. "Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies." Nature Genetics 47 (7): 702–9.

Purcell, Shaun M., Jennifer L. Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, et al. 2014. "A Polygenic Burden of Rare Disruptive Mutations in Schizophrenia." Nature 506 (7487): 185–90.

Rajji, T. K., Z. Ismail, and B. H. Mulsant. 2009. "Age at Onset and Cognition in Schizophrenia: Meta-Analysis." The British Journal of Psychiatry: The Journal of Mental Science 195 (4): 286– 93.

Robinson, Delbert G., Margaret G. Woerner, Marjorie McMeniman, Alan Mendelowitz, and Robert M. Bilder. 2004. "Symptomatic and Functional Recovery from a First Episode of Schizophrenia or Schizoaffective Disorder." The American Journal of Psychiatry 161 (3): 473– 79.

Ruderfer, Douglas M., Alexander W. Charney, Ben Readhead, Brian A. Kidd, Anna K. Kähler, Paul J. Kenny, Michael J. Keiser, et al. 2016. "Polygenic Overlap between Schizophrenia Risk and Antipsychotic Response: A Genomic Medicine Approach." The Lancet. Psychiatry 3 (4): 350–57.

Saha, Sukanta, David Chant, and John McGrath. 2007. "A Systematic Review of Mortality in Schizophrenia: Is the Differential Mortality Gap Worsening over Time?" Archives of General Psychiatry 64 (10): 1123–31.

Salomon, Joshua A., Juanita A. Haagsma, Adrian Davis, Charline Maertens de Noordhout, Suzanne Polinder, Arie H. Havelaar, Alessandro Cassini, et al. 2015. "Disability Weights for the Global Burden of Disease 2013 Study." The Lancet. Global Health 3 (11): e712–23.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." Nature.

Sham, P. C., C. J. MacLean, and K. S. Kendler. 1994. "A Typological Model of Schizophrenia Based on Age at Onset, Sex and Familial Morbidity." Acta Psychiatrica Scandinavica 89 (2): 135–41.

Singh, Tarjinder, James T. R. Walters, Mandy Johnstone, David Curtis, Jaana Suvisaari, Minna Torniainen, Elliott Rees, et al. 2017. "The Contribution of Rare Variants to Risk of Schizophrenia in Individuals with and without Intellectual Disability." Nature Genetics 49 (8): 1167–73.

So, Hon-Cheong, Carlos Kwan-Long Chau, Wan-To Chiu, Kin-Sang Ho, Cho-Pong Lo, Stephanie Ho-Yue Yim, and Pak-Chung Sham. 2017. "Analysis of Genome-Wide Association Data Highlights Candidates for Drug Repositioning in Psychiatry." Nature Neuroscience 20 (10): 1342–49.

Stefansson, Hreinn, Roel A. Ophoff, Stacy Steinberg, Ole A. Andreassen, Sven Cichon, Dan Rujescu, Thomas Werge, et al. 2009. "Common Variants Conferring Risk of Schizophrenia." Nature 460 (7256): 744–47.

Strassnig, M., J. Signorile, C. Gonzalez, and P. D. Harvey. 2014. "Physical Performance and Disability in Schizophrenia." Schizophrenia Research. Cognition 1 (2): 112–21.

Sugden, Karen, Eilis J. Hannon, Louise Arseneault, Daniel W. Belsky, Jonathan M. Broadbent, David L. Corcoran, Robert J. Hancox, et al. 2019. "Establishing a Generalized Polyepigenetic Biomarker for Tobacco Smoking." Translational Psychiatry 9 (1): 92.

Sullivan, Patrick F., Arpana Agrawal, Cynthia M. Bulik, Ole A. Andreassen, Anders D. Børglum, Gerome Breen, Sven Cichon, et al. 2017. "Psychiatric Genomics: An Update and an Agenda." The American Journal of Psychiatry, October.

Sullivan, Patrick F., Kenneth S. Kendler, and Michael C. Neale. 2003. "Schizophrenia as a Complex Trait: Evidence from a Meta-Analysis of Twin Studies." Archives of General Psychiatry 60 (12): 1187–92.

Szatkiewicz, J. P., C. O'Dushlaine, G. Chen, K. Chambert, J. L. Moran, B. M. Neale, M. Fromer, et al. 2014. "Copy Number Variation in Schizophrenia in Sweden." Molecular Psychiatry 19 (7): 762–73.

Taylor, Warren D., and Charles F. Reynolds. 2020. "Psychiatry's Obligation to Treat and Mitigate the Rising Burden of Age-Related Mental Disorders." JAMA Psychiatry.

Teschendorff, Andrew E., and Caroline L. Relton. 2018. "Statistical and Integrative System-Level Analysis of DNA Methylation Data." Nature Reviews. Genetics 19 (3): 129–47.

Torkamani, Ali, Nathan E. Wineinger, and Eric J. Topol. 2018. "The Personal and Clinical Utility of Polygenic Risk Scores." Nature Reviews. Genetics 19 (9): 581–90.

Turner, Adam W., Doris Wong, Caitlin N. Dreisbach, and Clint L. Miller. 2018. "GWAS Reveal Targets in Vessel Wall Pathways to Treat Coronary Artery Disease." Frontiers in Cardiovascular Medicine 5 (June): 72.

Waaijer, Mariëtte E. C., William E. Parish, Barbara H. Strongitharm, Diana van Heemst, Pieternella E. Slagboom, Anton J. M. de Craen, John M. Sedivy, Rudi G. J. Westendorp, David A. Gunn, and Andrea B. Maier. 2012. "The Number of p16INK4a Positive Cells in Human Skin Reflects Biological Age." Aging Cell 11 (4): 722–25.

Wahl, Simone, Alexander Drong, Benjamin Lehne, Marie Loh, William R. Scott, Sonja Kunze, Pei-Chien Tsai, et al. 2017. "Epigenome-Wide Association Study of Body Mass Index, and the Adverse Outcomes of Adiposity." Nature 541 (7635): 81–86.

Walsh, Tom, Jon M. McClellan, Shane E. McCarthy, Anjené M. Addington, Sarah B. Pierce, Greg M. Cooper, Alex S. Nord, et al. 2008. "Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia." Science 320 (5875): 539–43.

Wolf, Erika J., Mark W. Logue, Filomene G. Morrison, Elizabeth S. Wilcox, Annjanette Stone, Steven A. Schichman, Regina E. McGlinchey, William P. Milberg, and Mark W. Miller. 2019. "Posttraumatic Psychopathology and the Pace of the Epigenetic Clock: A Longitudinal Investigation." Psychological Medicine 49 (5): 791–800.

World Health Organization, WHO Staff, and WHO. 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organization.

Zhang, Yi, and Tatiana G. Kutateladze. 2018. "Diet and the Epigenome." Nature Communications.

PART 1

Functional investigations of schizophrenia biology

CHAPTER 3

A longitudinal model of human neuronal differentiation for functional investigation of schizophrenia polygenic risk

Authors
Anil P. S. Ori
Merel H. M. Bot
Remco T. Molenhuis
Loes M. Olde Loohuis
Roel A. Ophoff

Affiliations
[1.] Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California, USA
[2.] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA

Abstract

Common psychiatric disorders are characterized by complex disease architectures with many small genetic effects that contribute and complicate biological understanding of their etiology. There is therefore a pressing need for *in vitro* experimental systems that allow for interrogation of polygenic psychiatric disease risk to study the underlying biological mechanisms. We have developed an analytical framework that integrates genome-wide disease risk from genome-wide association studies with longitudinal *in vitro* gene expression profiles of human neuronal differentiation. We demonstrate that the cumulative impact of risk loci of specific psychiatric disorders is significantly associated with genes that are differentially expressed and upregulated during differentiation. We find the strongest evidence for schizophrenia, a finding that we replicate in an independent dataset. A longitudinal gene cluster involved in synaptic function primarily drives the association with schizophrenia risk. These findings reveal that *in vitro* human neuronal differentiation can be used to translate the polygenic architecture of schizophrenia to biologically relevant pathways that can be modeled in an experimental system. Overall, this work emphasizes the use of longitudinal *in vitro* transcriptomic signatures as a cellular readout and the application to the genetics of complex traits.

Manuscript status: published in Biological Psychiatry 2019 Apr 1;85(7):544-553
https://doi.org/10.1016/j.biopsych.2018.08.019

Introduction

Major psychiatric disorders feature a high heritability but have a largely unknown etiology(D. H. Geschwind and Flint 2015; Polderman et al. 2015). The increasing sample sizes of genome-wide association studies (GWAS) successfully result in identification of more susceptibility loci for these disorders(Sullivan et al. 2017). A major challenge is to understand and interpret the cumulative impact of many loci that collectively contribute to psychiatric disease risk and how to translate this complex polygenic architecture to biological pathways that drive the underlying molecular and cellular disease processes. Lack of applicable *in vitro* model systems and a framework to study polygenic psychiatric risk hinders the translation of genetics findings to disease biology(Falk et al. 2016).

Early brain development has been implicated in psychiatric disorders such as schizophrenia (SCZ)(Gulsuner et al. 2013; Purcell et al. 2014; Olde Loohuis et al. 2015; H. K. Finucane et al. 2015), autism spectrum disorder (ASD)(Daniel H. Geschwind 2011; Rubeis et al. 2014), and self-reported depression (SRD)(Hyde et al. 2016). Differentiation of human embryonic stem cells (hESCs) into neuronal lineages has been demonstrated to hold great promise to model early brain development(Shi et al. 2012; van de Leemput et al. 2014; Stein et al. 2014), and may thus offer a unique opportunity to study psychiatric disease biology *in vitro*. However, it has remained unclear whether the molecular dynamics underlying *in vitro* human neuronal differentiation are associated with polygenic psychiatric disease susceptibility.

We set out to investigate *in vitro* human neuronal differentiation in the context of polygenic psychiatric disease risk. To accomplish this, we performed a densely-sampled time series experiment and robustly detected transcriptome-wide changes across neuronal differentiation. To study the aggregate impact of risk loci, we integrated longitudinal *in vitro* gene expression signatures with GWAS summary statistics of major psychiatric disorders. We observe significant enrichment of genetic risk for multiple disorders in genes that are upregulated across differentiation. We further show that this effect is strongest for SCZ and primarily driven by a longitudinal gene cluster that is involved in synaptic functioning. These findings support to use of *in vitro* neuronal differentiation as a promising model system to study genetic psychiatric risk, particularly in the context of schizophrenia.

Methods

Approval for stem cell research

This study and all described work was approved by the University of California, Los Angeles Embryonic Stem Cell Research Oversight (ESCRO) committee.

In vitro human neuronal differentiation

WA09(H9)-derived hNSCs were commercially obtained (Gibco) as neural progenitors and subsequently expanded as adherent culture according to the manufacturer's guidelines. Low passage hNSCs (< 4 passage rounds) were plated in 12-well plates coated with poly-D-lysine (0.1 mg/mL, VWR) and laminin (4.52ug/cm2, CorningTM) at 1.5x105 cells, which were equally distributed and subsequently cultured in expansion medium as described above. After 24h of

proliferation, media was changed to neuronal differentiation medium consisting of Neurobasal® Medium (Gibco), 2% B-27® Serum-Free Supplement (Gibco), 2mM GlutaMax™-I Supplement, 0.05 mM β-mercaptoethanol (Gibco), and 1x Pen Strep. Media was changed every 2-3 days.

Experimental design and assessment of gene expression

Human neural stem cells were differentiated over a course of 30 days and RNA harvested at seven time points (day 0, 2, 5, 10, 15, 20, and 30) in triplicates or quadruplicates (n = 24). Genome-wide array-based transcriptome data was collected at the UCLA Neuroscience Genomics Core using Illumina's HumanHT-12 v4 Expression BeadChip Kit.

Data pre-processing and quality control

Gene expression data was extracted using the Gene Expression Module in GenomeStudio Software 2011.1. Data was background corrected with subsequent variance-stabilizing transformation and robust spline normalization was applied(Du, Kibbe, and Lin 2008; Lin et al. 2008). We excluded low quality probes and subsequently performed sample outlier detection by Euclidean distance and standardized connectivity. The FactoMineR package (v1.28) in R was used to perform principal component analysis (PCA). For subsequent downstream analyses, we used the normalized expression values of 19,012 high quality filtered probes for all 24 samples.

Transcriptome-based *in vitro* cellular identity

To investigate *in vitro* cellular identity across differentiation, we used transcriptomic signatures of cell-type specific genes of seven main cell types identified in the mouse cerebral cortex (Zhang et al. 2014). We extracted normalized gene expression values of these genes for each cell type from our own *in vitro* dataset and calculated mean standardized expression levels of cell type-specific genes for each of the seven cell types across days of differentiation.

Transition mapping to a spatiotemporal atlas of early human brain development

To investigate global transcriptomic matching between *in vitro* gene expression profiles and *in vivo* gene expression profiles of neocortical brain regions, we applied transition mapping (TMAP), which is implemented in the online CoNTExT bioinformatic pipeline (https://context.semel.ucla.edu)(Stein et al. 2014). Analyses were run for *in vitro* time points day-0 vs day-30, day-0 vs day-5, day-5 vs day-15, and day-15 vs day-30 across both temporal and spatial dimensions of human cortical development.

Time-series differential gene expression and cluster analysis

Two multivariate empirical Bayes models were used to identify differentially expressed genes across differentiation. We computed the one-sample T2-statistic and a probability of being differentially expressed using the mb.long() function in the Timecourse package (v 1.42) and the betr() function in the BETR package (v 1.26) in R, respectively(Tai and Speed 2006a), (Aryee et al. 2009). As both methods rank probes by their differential expression over time, differentially expressed genes were classified as the union of the set of probes with a probability of 1.0 using

BETR and an equally sized set of top ranked probes using the T2-statistic. We subsequently applied fuzzy c-means clustering to all differentially expressed probes and computed cluster membership values using the fclusList() and membership() function in the Mfuzz package in R(Kumar and E Futschik 2007; Schwämmle and Jensen 2010). Clusters were annotated using Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6.8) (Huang, Lempicki, and Sherman 2009) and probes with a membership > 0.5.

Integration of GWAS data with *in vitro* transcriptomic signatures

Illumina probe IDs were mapped to Ensembl gene IDs using NCBI build 37.3, duplicate IDs removed, and gene boundaries extended symmetrically by 10kb to include regulatory regions. Annotation files were then created mapping each gene ID or chromosomal position with *in vitro* gene parameters of interest, such as T2-statistic and cluster membership values. These files were then used as input to Multi-marker Analysis of GenoMic Annotation (MAGMA) and stratified LD score regression (sLDSR) to integrate *in vitro* signatures with GWAS data and study the cumulative impact across risk loci.

GWAS summary statistics and ancestry matched reference panels

GWAS summary statistics were obtained for SCZ(Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), major depressive disorder (MDD) (CONVERGE Consortium 2015), SRD(Hyde et al. 2016), bipolar disorder (BPD)(Group 2011), ASD(The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium 2017), attention deficit hyperactivity disorder (ADHD) (Demontis et al. 2017), cross disorder(Consortium 2013), Alzheimer's disease (AD)(Lambert et al. 2013), and adult human height (Wood AR Esko T Yang J Vedantam S Pers TH Gustafsson S Chu AY Estrada K Luan J Kutalik Z Amin N Buchkovich ML Croteau-Chonka DC Day FR Duan Y Fall T Fehrmann R Ferreira T Jackson AU Karjalainen J Lo KS Locke AE Mägi R Mihailov E Por 2014) (Supplemental Table S2). For each trait we used the most recent GWAS summary statistics that was publically available at the time of the analysis. The 1000 Genomes Project Phase 3 release (1KG) was used as reference panel to model ancestry-matched LD (1000 Genomes Project Consortium et al. 2015).

MAGMA gene-set analysis

MAGMA (v1.06)(de Leeuw et al. 2015) was used to perform gene-set analyses of GWAS data. MAGMA uses a multiple regression framework to associate a continuous or binary gene variable to GWAS gene level p-values. For each GWAS phenotype, we generated gene-level p-values by computing the mean SNP association using the default gene model ('snp-wise=mean') with +/- 10kb extensions of gene boundaries and SNPs with minor allele frequency (MAF) > 5%. For each annotation, we then regressed gene-level GWAS test statistics on the corresponding gene annotation variable using the '--gene-covar' function while adjusting for gene size, SNP density, and LD-induced correlations ('--model correct=all'), which is estimated from an ancestry-matched 1KG reference panel. Testing only for a positive association, i.e. enrichment of GWAS signal, we report one-sided p-values along with the corresponding regression coefficient.

Stratified LD Score Regression

We applied an extension to stratified LD score regression (sLDSR), a statistical method that partitions SNP-based heritability (h^2) from GWAS summary statistics (H. K. Finucane et al. 2015). This allows us to quantify the effects of continuous-valued annotations on the heritability (Gazal et al. 2017). For each annotation, we first estimated partitioned LD scores using the ldsc.py --l2 function with MAF > 5%, a 1 centimorgan (cm) window, and an ancestry-match 1KG reference panel. We ran sLDSR (ldsc.py --h^2) for each annotation of interest while accounting for the full baseline model, as recommended by the developers (H. K. Finucane et al. 2015; Gazal et al. 2017), and an extra annotation of all genes detected in our *in vitro* model (n = 12,414). As we only test for a positive association, we report the contribution to the per-SNP h^2 (τ) and the associated one-sided p-value, which is calculated using standard errors that are obtained via a block jackknife procedure (Bulik-Sullivan et al. 2015; H. K. Finucane et al. 2015).

Results

Longitudinal *in vitro* gene expression profiling confirms neuron-specific differentiation and matches *in vivo* human cortical development

To study the molecular dynamics underlying *in vitro* human neuronal differentiation, we differentiated an hNSC line (WA09/H9) to a neuronal lineage across 30 days. Genome-wide gene expression profiles were assayed densely at seven time points in at least triplicates (n=24 samples). To verify that the data was in agreement with the intended differentiation protocols, we investigated specific gene expression signatures over time. We first examined gene expression patterns of traditional gene markers (Tanapat 2013; Magavi and Macklis 2002) and found that neural stem cell and proliferation markers (MKI67, Nestin, and SOX2) are downregulated, while early neuronal markers (BDNF and DCX) are upregulated as differentiation progresses (Figure 1A-B). MAP2, a more mature neuronal marker(Tanapat 2013; von Bohlen Und Halbach 2007), is first upregulated and subsequently downregulated at later time points, suggesting that the differentiated culture maintains a relatively immature neuronal identity.

Next, we explored PCA on normalized gene expression values using the full transcriptome and found a large proportion of the variance in expression to be explained by the differentiation process, with minimal effects of technical variation (Figure 1C & S1). Investigation of transcriptome-based cell type-specific gene expression signatures of major classes of cell types in the cerebral cortex shows that relative neuronal gene expression increases as neuronal differentiation progresses over time (Figure 1D). There is no evidence of glial- or endothelial-specific gene expression, which confirms a broadly neuronal *in vitro* cellular identity.

Having established that the *in vitro* differentiation process is predominantly neuronal, we applied transition mapping (TMAP) to assess the correspondence of longitudinal *in vitro* transcriptome data to *in vivo* signatures of both brain developmental stages and laminae of the human neocortex. We find significant matching between the *in vitro* longitudinal DGE profiles (day-0 vs day-30) and *in vivo* developmental stage from 4 weeks post-conception (PCW) to 24 PCW (Figure S2).

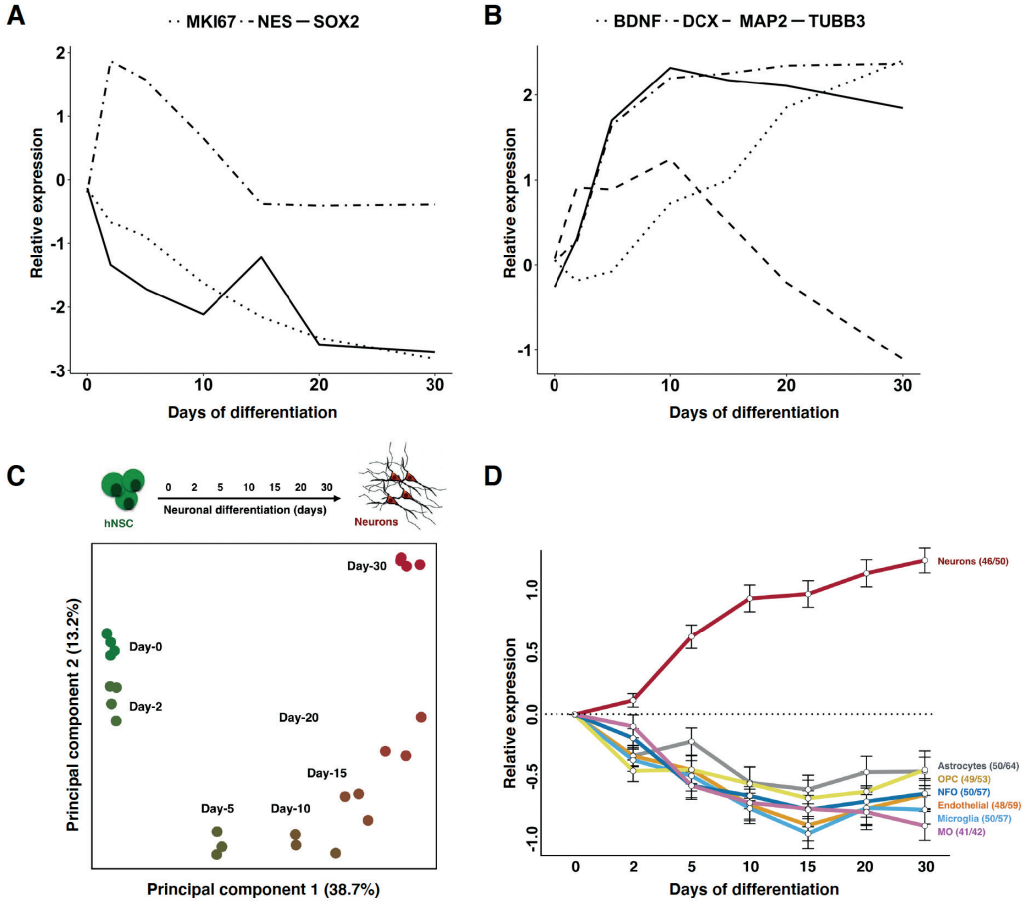


Figure 1. In vitro gene expression profiles confirm a neuron-specific differentiation process. *Relative gene expression of traditional stem cell (A) and neuronal (B) markers plotted across days of differentiation. (C) PCA of in vitro transcriptomic data with PC1 (x-axis) and PC2 (y-axis) visualized. Variance explained per component is shown in parentheses. (D) Transcriptome-based cellular identity is shown by average expression of cell type specific genes across days of differentiation. The first number in the parentheses represents the number of genes for which the average expression is plotted. The second number represents the corresponding number of probes assayed. OPC = oligodendrocyte precursor cells, NFO = newly formed oligodendrocytes, MP = myelinating oligodendrocytes.*

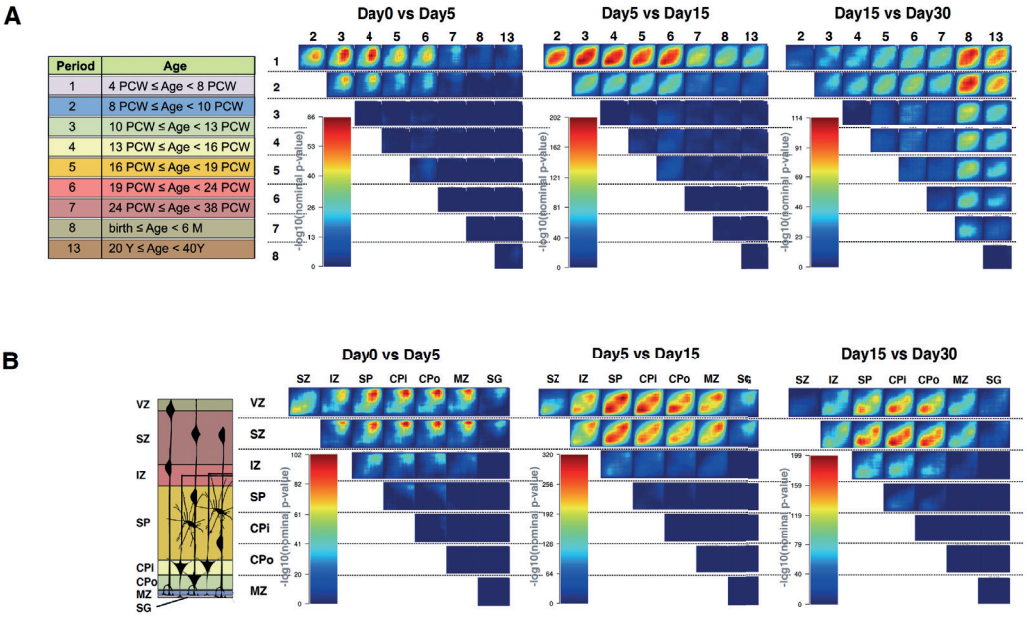


Figure 2. In vitro gene expression profiles match in vivo human cortical development. *TMAP output visualizes the amount of overlap between in vitro and in vivo DGE profiles colored by –log10(p-value) (see figure S2 for more details on interpretation). Note that p-values are shown on varying color scales between graphs. Abbreviations and numbering above maps correspond to schematic representations on the left (adopted from Stein et al., 2014) of different developmental stages (A) and laminae (B). VZ = ventricular zone, SZ = subventricular zone, IZ = intermediate zone, SP=subplate zone, CPi= inner cortical plate, CPo = outer cortical plate, MZ = marginal zone, PCW = post conception weeks, M = months, Y = years, Period = developmental stage.*

This overlaps with the primary period of neurogenesis in the neocortex, which starts around 6 PCW (Clancy, Darlington, and Finlay 2001; Stiles and Jernigan 2010). To gain more insight into this overlap, we partitioned the TMAP analyses in three comparisons and examined how *in vitro* to *in vivo* matching progressed over time across differentiation. We see a clear progression in matching from early developmental stages to later stages (Figure 2A). For example, *in vitro* day-0 vs day-5 show strong overlap with *in vivo* period-1 (4-8 PCW) vs period-4 (13-16 PCW), while *in vitro* day-15 vs day-30 shows stronger overlap with *in vivo* period-2 (8-10 PCW) vs period-8 (birth-6M). Similarly, *in vitro* longitudinal DGE shows progression from overlap of early time points with inner laminae, to overlap with more upper cortical layers as *in vitro* neuronal differentiation advances (Figure 2B and S2).

In vitro neuronal differentiation reveals specific longitudinal gene clusters

To identify biological pathways associated with neuronal differentiation, we applied an analysis framework specifically tailored to time-series gene expression data (see Methods and Supplemental Methods). A total of 7,734 probes, mapping to 5,818 genes, were differentially expressed over time (Figure S3). We find that these genes are, on average, more constrained to genetic variation compared to non-differentially expressed genes (section S2). Using only differentially expressed probes, we next applied fuzzy c-means clustering and identified eight distinct longitudinal gene clusters (Figure 3 and S4). For each probe, we generated a corresponding cluster membership value, representing the degree to which a gene belongs to a cluster. To identify most informative biological interpretation of each cluster, we analyzed genes with high cluster membership for enrichment of functional annotations using DAVID (Supplemental Methods and Table S1).

We identified three clusters with decreasing gene expression over time that are significantly enriched for cell division and RNA regulation and processing genes, reflective of stem cell proliferation and cell fate determination that is tightly controlled and regulated by RNA dependent processes (Hattori, Buac, and Ito 2016). Second, there are three clusters showing increased gene expression levels over time that are primarily enriched for neuronal processes, such as neuron formation and synaptic function. Another independent cluster shows an inverted U-shaped expression pattern during development, enriched for genes involved in transcriptional regulation. The final cluster is enriched for genes involved in extracellular region and cell adhesions. These processes are important for cell connectivity and have also been implicated in cell proliferation and neuronal migration (Barros, Franco, and Muller 2011; Bikbaev, Frischknecht, and Heine 2015). Together, these eight gene clusters reveal different biological mechanisms that are associated with neuronal differentiation and consistent with known biology of neurodevelopment. We hypothesize that the study of these longitudinal gene expression clusters can help decipher disease mechanisms involved in psychiatric phenotypes.

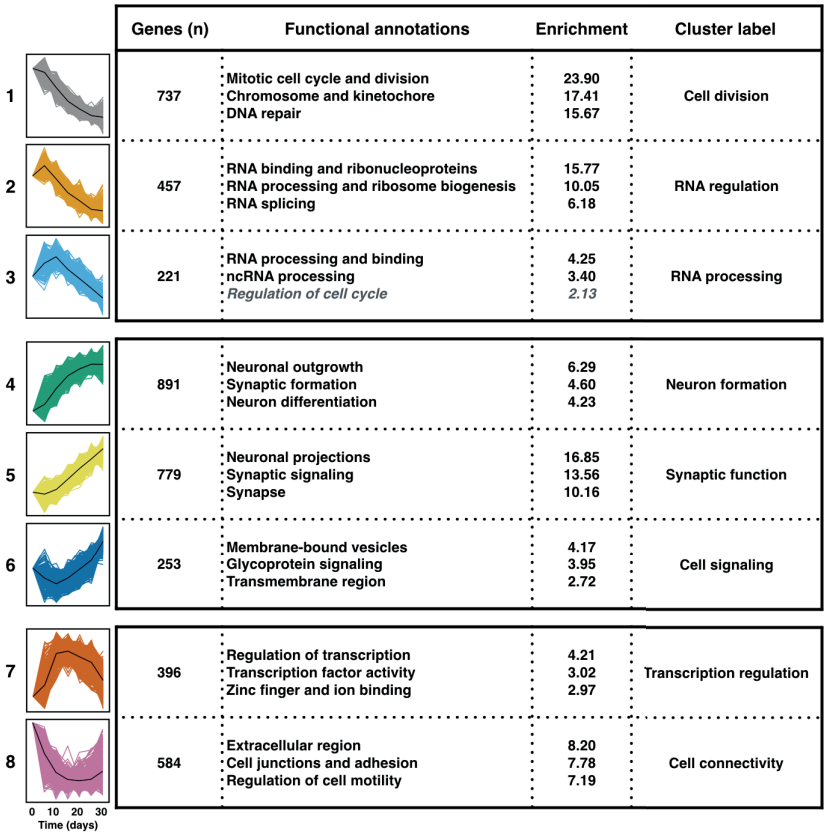


Figure 3. Identified gene clusters highlight biological pathways important for neuronal differentiation. Top significant functional annotations and corresponding enrichment score are shown for each gene cluster. Longitudinal gene expression is visualized for high member genes only (black line represents mean gene expression). Each cluster is color-coded with the number of genes at membership > 0.5 denoted. See table S1 for full annotation results.

Differentially expressed genes are enriched for polygenic psychiatric disease risk

To examine how aggregate psychiatric disease risk is distributed across genes that are important for neuronal differentiation, we applied gene-set analysis and partitioning of h^2 with MAGMA and sLDSR, respectively. We used GWAS summary statistics from major psychiatric disorders in addition to Alzheimer's disease (AD) and adult human height, which served as non-psychiatric control phenotypes that are heritable and polygenic. Using a two-step approach, we first investigated disease susceptibility on overall differential expression level and subsequently proceeded to deconstruct these associations across the longitudinal gene clusters. We find that genes that are differentially expressed are enriched for genetic risk of multiple psychiatric disorders. We find significant effects with MAGMA for SCZ (P=0.001), ADHD (P=0.002), and SRD (P=0.003) (Table 1 and Table S3). With sLDSR, we find nominally significant effects for SCZ (P=0.01)

and SRD (P=0.02) and a suggestive association for ADHD (P=0.06) (Table 1 and Table S4). We observed a suggestive enrichment for BPD, and no enrichment for the cross disorder, ASD, MDD CONVERGE or for adult height and AD.

	MAGMA			LDSC	
Phenotype	**Beta (SE)**	**Beta_std**	**P-value**	**per-SNP h^2 (SE)**	**P-value**
Psychiatric					
Schizophrenia	0.022 (0.0065)	0.097	**8.66 x 10-4**	9.36×10^{-9} (4.13×10^{-9})	0.02
MDD PGC	0.009 (0.0044)	0.043	0.03	1.30×10^{-8} (1.22×10^{-8})	0.29
ADHD	0.008 (0.0042)	0.036	0.06	2.08×10^{-8} (4.04×10^{-8})	0.60
Bipolar disorder	0.008 (0.0054)	0.036	0.13	1.43×10^{-8} (6.83×10^{-8})	0.04
MDD CONVERGE	0.005 (0.0042)	0.021	0.27	7.03×10^{-8} (2.52×10^{-8})	**5.36×10^{-3}**
ASD	-0.003 (0.0041)	-0.015	0.41	9.56×10^{-9} (2.00×10^{-8})	0.63
Cross disorder	0.003 (0.0048)	0.014	0.51	4.48×10^{-9} (4.66×10^{-9})	0.34
Neurodegenerative					
Alzheimer's disease	0.003 (0.0044)	0.014	0.48	-2.46×10^{-10} (3.99×10^{-9})	0.95
Non-brain					
Height*	-0.003 (0.0107)	-0.012	0.79	-1.20×10^{-8} (5.50×10^{-9})	0.03

Table 1. Differentially expressed genes are enriched for polygenic risk of multiple psychiatric disorders. *Shown are results of MAGMA and sLDSR for differentially expressed genes. P-values highlighted in bold show phenotypes that survive multiple testing correction (n=9). See Table S3 and S4 for more details. Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given a change of one standard deviation in log T2 statistic, τ(tau) = the contribution to the per-SNP h^2.*

We next investigated whether enrichment across differentially expressed genes was driven by up- or downregulation of genes during differentiation. For SCZ, we find that the effect is driven by genes that are upregulated (MAGMA P=5.0x10-7, sLDSR P=6.1x10-5) and not by genes that are downregulated (MAGMA P=0.98, sLDSR P=0.61) (Figure 4 and Figure S6). For SRD, we only find a stronger enrichment in upregulated genes with MAGMA (P=3.5x10-4), while ADHD shows no specific evidence for either up or downregulated genes.

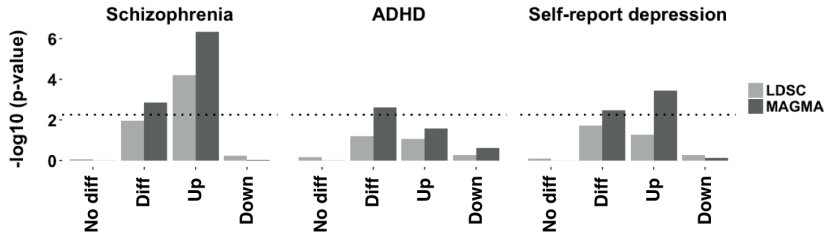


Figure 4. Schizophrenia polygenic risk lies in genes up-regulated during neuronal differentiation. *A more detailed investigation of the effect of differentially expressed genes on the heritability of SCZ, ADHD, and SRD. The y-axis denotes the –log10 P-value of the enrichment. No diff = genes that are not differentially expressed; Diff = log (T2-statistic) as shown in Table 1; Up = genes upregulated during differentiation; Down = genes downregulated during differentiation. The dotted line represents the threshold for P=0.0056 (n=9 traits).*

Psychiatric disease risk aggregates to specific longitudinal gene clusters

Next, we explored the relationship between differentially expressed genes and disease risk on cluster level. For this analysis, we only included traits that show significant disease enrichment across differentially expressed genes using MAGMA after correcting for multiple testing (SCZ, ADHD, SRD) and our control traits (AD, height). These disease traits showed at least a nominally significant effect with sLDSR as well. Using both MAGMA and sLDSR, we integrated cluster membership values with GWAS summary statistics (n=5) and assessed whether genome-wide disease risk aggregates to any of the eight experimentally identified longitudinal gene clusters. Overall, MAGMA and sLDSR show a strong concordance across phenotypes and clusters (rho = 0.92, p<2.2x10-16, n=40, see also Figure S7). After Bonferroni correction (n=40), we find five significant phenotype-cluster associations with MAGMA and three with sLDSR (Figure 5 and Table S5/S6).

We find that multiple upregulated clusters show enrichment for SCZ with the strongest evidence for the synaptic function cluster (MAGMA P=1.8x10-7, sLDSR P=7.2x10-5) (see Figure S8). For SRD, we find significant associations in the transcription regulation (P=2.5x10-5) and the neuron formation (P=1.2x10-4) gene cluster with MAGMA only. While the analysis of adult height using all differentially expressed genes did not yield any evidence for enrichment of genetic signal, enrichment is observed at the cluster level. The cell connectivity cluster (P=3.7x10-4) is enriched for height, in addition to suggestive enrichments in the cell division and RNA regulation cluster, which are not present for any of the psychiatric phenotypes. Remarkably, across all 8 clusters the enrichments of SCZ and height are inversely correlated (rho=-0.85, P=0.011, n=8; see also section S3 and Figure S9-10).

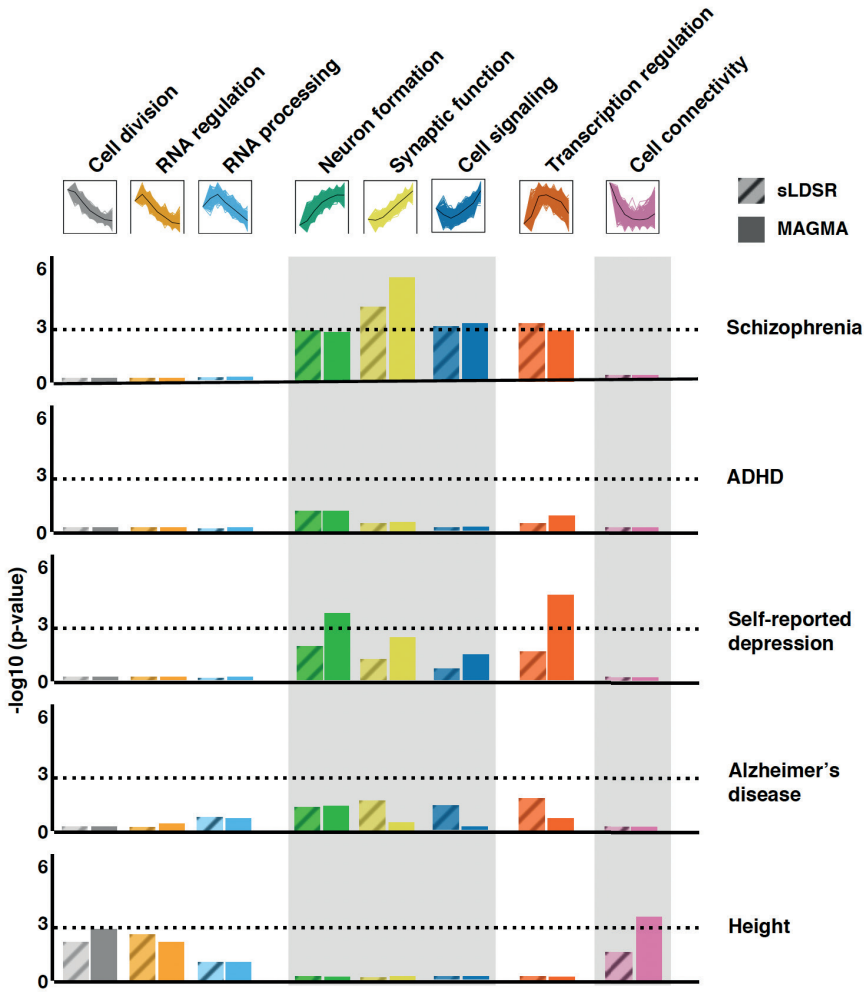


Figure 5. Polygenic psychiatric risk is distributed across specific longitudinal gene clusters. *Results from sLDSC (diagonal pattern) and MAGMA (solid colors) are shown for each phenotype (labels on the right) colored by gene cluster. Gene cluster annotation and cluster expression pattern are shown on top. The y-axis states the –log10 (p-value). The dotted horizontal line represents the threshold for Bonferroni correction (p=0.05/40).*

Finally, in order to take into account the full spectrum of correlations and dependencies between clusters (Figure S11), we performed a conditional analysis for SCZ, the trait for which the strongest cluster enrichments are observed with both methods. Using the same MAGMA model, for each cluster, we conditioned on the highest gene members (membership > 0.5) of the other seven clusters (Table 2). We find that the SCZ enrichment is driven by the synaptic function

cluster (p=2.88x10-3) only. The same conditional analysis for SRD, which only showed a significant enrichment with MAGMA, shows that this effect is primarily driven by the transcription regulation cluster (p=5.42x10-3) (Table S7).

	MAGMA Primary		MAGMA Conditional	
Schizophrenia - clusters	**Beta (SE)**	**P-value**	**Beta (SE)**	**P-value**
Cell division	-0.045 (0.017)	1.00	-0.047 (0.027)	0.96
RNA regulation	-0.040 (0.017)	0.99	-0.044 (0.027)	0.95
RNA processing	-0.006 (0.017)	0.64	-0.011 (0.024)	0.68
Neuron formation	0.048 (0.017)	$2.12x10^{-3}$	0.018 (0.036)	0.30
Synaptic function	**0.077 (0.017)**	**$1.82x10^{-6}$**	**0.070 (0.026)**	**$2.88x10^{-3}$**
Cell signaling	0.052 (0.016)	$6.88x10^{-4}$	0.032 (0.023)	0.08
Transcription regulation	0.048 (0.016)	$1.67x10^{-3}$	0.019 (0.025)	0.22
Cell connectivity	-0.061 (0.017)	1.00	-0.076 (0.026)	1.00

Table 2. The association with SCZ risk is driven by the synaptic function gene cluster. *Gene level association signal is regressed on cluster membership while adjusting for high membership genes of all other seven clusters. Shown are the results of the primary analysis (not adjusted for other clusters) and the conditional analysis with MAGMA. Beta = regression coefficient, SE = standard error.*

Replication in the CORTECON RNA-seq dataset shows strong concordance with discovery analyses

To evaluate reproducibility of our findings, we performed a comprehensive replication analysis in the CORTECON RNA sequencing (RNA-seq) dataset of *in vitro* human cortical differentiation (van de Leemput et al. 2014). While the CORTECON project was executed using widely different experimental procedures (section S4.1), we detect largely overlapping transcriptomic patterns with the discovery dataset. Between datasets, we see robust sample correlations across the differentiation trajectory (section S4.2, Figure S12), including in stem cell and early neuronal gene marker expression patterns (section S4.4, Figure S14-15). We observe a highly significant overlap in differentially expressed genes (section S4.5) and in identified gene clusters (section S4.6, Figure S16-17). We in addition find that genes differentially expressed during 37 days of differentiation in CORTECON, which closely maps to 30 days of differentiation in the discovery set, are significantly associated with SCZ risk (beta=0.047, P=0.007, section S4.7). As in the discovery dataset, this association is driven by genes that are upregulated over time (P=0.008) but not downregulated (P=0.74). While the identified gene clusters show significant overlap with the eight gene clusters from the discovery analysis (Figure S17), we do not observe the association with SCZ risk to be distributed to a single gene cluster.

To investigate whether similar genes are driving the association with SCZ risk between our discovery analysis and the CORTECON dataset, we adjusted our analysis in the CORTECON dataset for the synaptic gene cluster (n= 779 genes) of the discovery analysis. We find that the strength of the association between SCZ risk and day-37 upregulated genes decreases when we account for synaptic genes from the discovery analysis (beta=0.044, P=0.031, section S4.7). We have highlighted a set of genes that have high membership to the synaptic gene cluster, are

differentially expressed in CORTECON, and are significantly associated to SCZ based on the GWAS (Figure S18). Taken together, this suggests that the same group of genes underlie the association between SCZ polygenic risk and transcriptomic signatures across differentiation and further demonstrates the concordance between both datasets.

Discussion

We investigated a longitudinal *in vitro* stem cell model of human neuronal differentiation to study psychiatric disease susceptibility based on evidence from GWAS. We confirmed that our *in vitro* model highlights transcriptomic profiles that are in line with an emerging neuronal identity that recapitulates signatures of *in vivo* cortical development across specific developmental time periods and laminae of the human neocortex. This is in line with previous findings (Stein et al. 2014) and highlights that longitudinal gene expression dynamics underlying our model of human neuronal differentiation can be informative to study genes and pathways involved in *in vivo* human cortical development. Importantly, neuronal cell types (Skene et al. 2017; Genovese et al. 2016; Forrest et al. 2017) and early brain development (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Olde Loohuis et al. 2015; H. Finucane et al. 2017) have been postulated as integral components of SCZ disease susceptibility. Here, we observe that genes differentially expressed across neuronal differentiation are significantly associated with genome-wide disease risk of SCZ, a finding that we replicate in an independent dataset. Our findings suggest that SCZ risk aggregates to genes involved in synaptic functioning during development. Although not the only pathogenic process contributing to SCZ, synaptic dysfunction is most strongly supported by genetic data, postmortem expression studies, and animal models (Genovese et al. 2016; Hall et al. 2015; Lips et al. 2011; Pocklington, O'Donovan, and Owen 2014; Schwarz et al. 2016; O'Dushlaine et al. 2015). We are the first to provide evidence for this hypothesis using a longitudinal *in vitro* cell-based model and aggregate polygenic disease risk. Our results suggest that high gene members of the synaptic function gene cluster enriched for SCZ (Figure S18), such as Calcium Voltage-Gated Channel Subunit Alpha 1C (CACNA1C), located at a genome-wide significant SCZ locus (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), are suitable candidates for functional follow-up in this *in vitro* model. We find no evidence for AD, a late-onset non-psychiatric brain disease, nor for adult human height in this neuronal cluster. Together, our findings demonstrate that longitudinal transcriptomic signatures important for neuronal differentiation recapitulate the *in vivo* context and align with the genetic basis of the disease. SCZ disease biology, and in particular synaptic functioning, can thus be studied through these molecular processes captured by this *in vitro* model.

We also observed a significant enrichment of genetic signal with MAGMA for SRD in genes upregulated during differentiation and show that this enrichment is predominantly driven by genes in the transcription regulation gene cluster. Interestingly, the SRD GWAS reported that the top SNPs were enriched for transcription regulation related to neurodevelopment (Hyde et al. 2016), which is in line with our *in vitro* findings. We observed no enrichment of the GWAS of recurrent and severe MDD in Han-Chinese women (CONVERGE Consortium 2015). The latter sample represents the most genetically and phenotypically homogeneous GWAS of MDD. The fact that for these results no enrichment for any of our gene sets was observed may suggest

that neurodevelopmental processes play a lesser role in MDD (Peterson et al. 2017). Alternatively, larger sample sizes are needed to better capture the genome-wide genetic risk associated with MDD (Figure S19). Self-reported depression is a much broader phenotype that may include other psychiatric traits, which could drive the observed neurodevelopment and transcription findings. Although it remains unclear how these results and the application of the model extrapolate to the MDD phenotype, our approach does highlight enrichment in distinct clusters for SRD and SCZ and could help shed light on how these two complex traits differ in their etiology.

A strength of our approach is the longitudinal analysis framework that we developed. We implemented an experimental design across a dense and repeatedly sampled time-series and integrated longitudinal transcriptomic signatures with genome-wide disease risk using available GWAS summary statistics. This increases statistical power to directly investigate the cumulative impact of risk loci on genes important to our model system. While we specifically chose to perform our experiments across an isogenic background to minimize variation and maximize statistical power to identify transcriptomic signatures, our framework can easily be extended to a multi-sample design (e.g. cases vs controls) (Tai and Speed 2006b; Aryee et al. 2009), which makes it relevant for many disease-specific experimental settings.

Our experimental procedure applied differentiation towards a broad neuronal phenotype. Our work does not exclude disease associations with specific subtypes of neuronal cells or other major brain cell types, nor does it exclude cell non-autonomous changes that may contribute. We provide a proof-of-concept of an *in vitro* model of neuronal cells for studying complex diseases, such as SCZ, and present an analytical framework that includes longitudinal assessment of gene expression profiles. This approach can readily be extended to study *in vitro* differentiation of other major brain cell types, such as astrocytes or oligodendrocytes. In addition, co-culture with astrocyte may facilitate a more mature neuronal culture (Tang et al. 2013; Johnson et al. 2007) and provide further insights into the temporal specificity of SCZ genetic risk. Although we show strong evidence for SCZ risk in early prenatal neurodevelopment, our findings do not preclude an additional contribution of postnatal neurodevelopment to the etiology of the disease (Birnbaum et al. 2014; Pers et al. 2015; Sekar et al. 2016).

In summary, as GWAS risk loci have small effect sizes and are abundantly distributed across the genome, new approaches are needed that allow for functional investigation of polygenic disease architectures. Embracing the polygenic nature of psychiatric disorders is an important step forward in translating findings from GWAS to disease biology. Our approach allowed us to narrow down on potential core disease processes and opens up new avenues to study disease in the context of polygenicity. Future work may for example incorporate model perturbations to study aggregate disease risk in finer detail or use the model for functional fine mapping of specific SCZ GWAS loci across an isogenic background in a controlled environment. Overall, this work contributes to understanding the functional mechanisms that underlie psychiatric disease heritability and polygenicity in the post GWAS era.

Declarations

Author Contribution

The project was led by R.O. Experiments were designed and conceived by A.O. and R.O. Experiments were optimized, conducted and samples processed by A.O., M.B., and R.M. Analysis of the data was performed by A.O and M.B and feedback provided by L.O. and R.O. The main findings were interpreted by A.O., M.B., R.M., L.O., and R.O. Primary drafting of the manuscript was performed by A.O. and main feedback provided by R.O and L.O. All authors contributed to the production and approval of the final manuscript.

Conflict of Interest

The authors report no biomedical financial interests or potential conflicts of interest.

Acknowledgement

We thank all research participants and researchers involved in making each GWAS summary statistic available and this work possible, including the 23andMe Research Team. We thank C. de Leeuw for his helpful input and troubleshooting with MAGMA analyses and thank the LD score regression team for their input and helpful troubleshooting with stratified LDSR. This research was supported by NIH/NIMH R01 MH090553 and U01MH105578.

Data availability

The Illumina HT-12 v4 gene expression data is available through the Gene Expression Omnibus (GEO) archive (GSE92845). This dataset has the raw and normalized gene expression values on all samples. Supplementary table 8 furthermore has specific probe annotations, such as probabilities of differential expression and probe membership values for all identified clusters.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74.

Aryee, Martin J., José A. Gutiérrez-Pabello, Igor Kramnik, Tapabrata Maiti, and John Quackenbush. 2009. "An Improved Empirical Bayes Approach to Estimating Differential Gene Expression in Microarray Time-Course Data: BETR (Bayesian Estimation of Temporal Regulation)." BMC Bioinformatics 10 (1): 409.

Barros, Claudia S., Santos J. Franco, and Ulrich Muller. 2011. "Extracellular Matrix: Functions in the Nervous System." Cold Spring Harbor Perspectives in Biology 3 (1): 1–24.

Bikbaev, Arthur, Renato Frischknecht, and Martin Heine. 2015. "Brain Extracellular Matrix Retains Connectivity in Neuronal Networks." Scientific Reports 5: 14527.

Birnbaum, Rebecca, Andrew E. Jaffe, Thomas M. Hyde, Joel E. Kleinman, and Daniel R. Weinberger. 2014. "Prenatal Expression Patterns of Genes Associated with Neuropsychiatric Disorders." The American Journal of Psychiatry 171 (7): 758–67.

Bohlen Und Halbach, O. von. 2007. "Immunohistological Markers for Staging Neurogenesis in Adult Hippocampus." Cell and Tissue Research 329 (3): 409–20.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of The Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." Nature Genetics 47 (3): 291–95.

Clancy, B., R. B. Darlington, and B. L. Finlay. 2001. "Translating Developmental Time across Mammalian Species." Neuroscience 105 (1): 7–17.

Consortium, Cross-Disorder Group of The Psychiatric Genomics. 2013. "Identification of Risk Loci with Shared Effects on Five Major Psychiatric Disorders: A Genome-Wide Analysis." The Lancet 381 (9875): 1371–79.

CONVERGE Consortium. 2015. "Sparse Whole-Genome Sequencing Identifies Two Loci for Major Depressive Disorder." Nature 523 (7562): 588.

Demontis, Ditte, Raymond K. Walters, Joanna Martin, Manuel Mattheisen, Thomas D. Als, Esben Agerbo, Rich Belliveau, et al. 2017. "Discovery of the First Genome-Wide Significant Risk Loci for ADHD." bioRxiv. https://doi.org/https://doi.org/10.1101/145581.

Du, Pan, Warren a. Kibbe, and Simon M. Lin. 2008. "Lumi: A Pipeline for Processing Illumina Microarray." Bioinformatics 24 (13): 1547–48.

Falk, A., V. M. Heine, A. J. Harwood, P. F. Sullivan, M. Peitz, O. Brüstle, S. Shen, et al. 2016. "Modeling Psychiatric Disorders: From Genomic Findings to Cellular Phenotypes." Molecular Psychiatry, no. vember 2015: 1167–79.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. "Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics." Nature Genetics 47 (11): 1228–35.

Finucane, Hilary, Yakir Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2017. "Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types." bioRxiv. https://doi.org/https://doi.org/10.1101/103069.

Forrest, Marc P., Hanwen Zhang, Winton Moy, Heather McGowan, Catherine Leites, Leonardo E. Dionisio, Zihui Xu, et al. 2017. "Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci." Cell Stem Cell 21 (3): 305–18.e8.

Gazal, Steven, Hilary K. Finucane, Nicholas A. Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, et al. 2017. "Linkage Disequilibrium–dependent Architecture of Human Complex Traits Shows Action of Negative Selection." Nature Genetics.

Genovese, Giulio, Menachem Fromer, Eli A. Stahl, Douglas M. Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L. Moran, et al. 2016. "Increased Burden of Ultra-Rare Protein-Altering Variants among 4,877 Individuals with Schizophrenia." Nature Neuroscience 19 (October): 1433–41.

Geschwind, Daniel H. 2011. "Genetics of Autism Spectrum Disorders." Trends in Cognitive Sciences.

Geschwind, D. H., and J. Flint. 2015. "Genetics and Genomics of Psychiatric Disease." Science 349 (6255): 1489–94.

Group, Psychiatric Gwas Consortium Bipolar Disorder Working. 2011. "Large-Scale Genome-Wide Association Analysis of Bipolar Disorder Identifies a New Susceptibility Locus near ODZ4." Nature Genetics 43 (10): 977–83.

Gulsuner, Suleyman, Tom Walsh, Amanda C. Watts, Ming K. Lee, Anne M. Thornton, Silvia Casadei, Caitlin Rippey, et al. 2013. "Spatial and Temporal Mapping of de Novo Mutations in Schizophrenia to a Fetal Prefrontal Cortical Network." Cell 154 (3): 518–29.

Hall, Jeremy, Simon Trent, Kerrie L. Thomas, Michael C. O'Donovan, and Michael J. Owen. 2015. "Genetic Risk for Schizophrenia: Convergence on Synaptic Pathways Involved in Plasticity." Biological Psychiatry.

Hattori, Ayuna, Kristina Buac, and Takahiro Ito. 2016. "Regulation of Stem Cell Self-Renewal and Oncogenesis by RNA-Binding Proteins." In RNA Processing Disease and Genome-Wide Probing, 153–88.

Huang, Da Wei, Richard a. Lempicki, and Brad T. Sherman. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." Nature Protocols 4 (1): 44–57.

Hyde, Craig L., Michael W. Nagle, Chao Tian, Xing Chen, Sara A. Paciga, Jens R. Wendland, Joyce Y. Tung, David A. Hinds, Roy H. Perlis, and Ashley R. Winslow. 2016. "Identification of 15 Genetic Loci Associated with Risk of Major Depression in Individuals of European Descent." Nature Publishing Group 48 (August): 1031–36.

Johnson, M. Austin, Jason P. Weick, Robert A. Pearce, and Su-Chun Zhang. 2007. "Functional Neural Development from Human Embryonic Stem Cells: Accelerated Synaptic Activity via Astrocyte Coculture." The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 27 (12): 3069–77.

Kumar, Lokesh, and Matthias E Futschik. 2007. "Mfuzz: A Software Package for Soft Clustering of Microarray Data." Bioinformation 2 (1): 5–7.

Lambert, J. C., C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, A. L. DeStafano, et al. 2013. "Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's Disease." Nature Genetics 45 (12): 1452–58.

Leemput, Joyce van de, Nathan C. Boles, Thomas R. Kiehl, Barbara Corneo, Patty Lederman, Vilas Menon, Changkyu Lee, et al. 2014. "CORTECON: A Temporal Transcriptome Analysis of *in Vitro* Human Cerebral Cortex Development from Human Embryonic Stem Cells." Neuron 83 (1): 51–68.

Leeuw, Christiaan A. de, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. 2015. "MAGMA: Generalized Gene-Set Analysis of GWAS Data." PLoS Computational Biology 11 (4).

Lin, Simon M., Pan Du, Wolfgang Huber, and Warren a. Kibbe. 2008. "Model-Based Variance-Stabilizing Transformation for Illumina Microarray Data." Nucleic Acids Research 36 (2): 1–9. Lips, E. S., L. N. Cornelisse, R. F. Toonen, J. L. Min, C. M. Hultman, P. a. Holmans, M. C. O'Donovan, et al. 2011. "Functional Gene Group Analysis Identifies Synaptic Gene Groups as Risk Factor for Schizophrenia." Molecular Psychiatry 4: 1–11.

Magavi, Sanjay S. P., and Jeffrey D. Macklis. 2002. "Immunocytochemical Analysis of Neuronal Differentiation." Methods in Molecular Biology 198: 291–97.

O'Dushlaine, Colm, Lizzy Rossin, Phil H. Lee, Laramie Duncan, Neelroop N. Parikshak, Stephen Newhouse, Stephan Ripke, et al. 2015. "Psychiatric Genome-Wide Association Study Analyses Implicate Neuronal, Immune and Histone Pathways." Nature Neuroscience 18 (2): 199–209.

Olde Loohuis, Loes M. Olde, Jacob A. S. Vorstman, Anil P. Ori, Kim A. Staats, Tina Wang, Alexander L. Richards, Ganna Leonenko, et al. 2015. "Genome-Wide Burden of Deleterious Coding Variants Increased in Schizophrenia." Nature Communications 6: 7501.

Pers, Tune H., Pascal Timshel, Stephan Ripke, Samantha Lent, Patrick F. Sullivan, Michael C. O'Donovan, Lude Franke, and Joel N. Hirschhorn. 2015. "Comprehensive Analysis of Schizophrenia-Associated Loci Highlights Ion Channel Pathways and Biologically Plausible Candidate Causal Genes." Human Molecular Genetics 25 (6): 1247–54.

Peterson, Roseann E., Na Cai, Tim B. Bigdeli, Yihan Li, Mark Reimers, Anna Nikulova, Bradley T. Webb, et al. 2017. "The Genetic Architecture of Major Depressive Disorder in Han Chinese Women." JAMA Psychiatry 74 (2): 162–68.

Pocklington, Andrew J., Michael O'Donovan, and Michael J. Owen. 2014. "The Synapse in Schizophrenia." The European Journal of Neuroscience 39 (7): 1059–67.

Polderman, Tinca J. C., Beben Benyamin, Christiaan A. de Leeuw, Patrick F. Sullivan, Arjen van Bochoven, Peter M. Visscher, and Danielle Posthuma. 2015. "Meta-Analysis of the Heritability of Human Traits Based on Fifty Years of Twin Studies." Nature Genetics 47 (7): 702–9.

Purcell, Shaun M., Jennifer L. Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, et al. 2014. "A Polygenic Burden of Rare Disruptive Mutations in Schizophrenia." Nature 506 (7487): 185–90.

Rubeis, Silvia De, Xin He, Arthur P. Goldberg, Christopher S. Poultney, and Kaitlin Samocha. 2014. "Synaptic, Transcriptional, and Chromatin Genes Disrupted in Autism A." Nature 515 (7526): 209–15.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." Nature 511(7510):421-7.

Schwämmle, Veit, and Ole Nørregaard Jensen. 2010. "A Simple and Fast Method to Determine the Parameters for Fuzzy c-Means Cluster Analysis." Bioinformatics 26 (22): 2841–48.

Schwarz, E., R. Izmailov, P. Lio, and A. Meyer-Lindenberg. 2016. "Protein Interaction Networks Link Schizophrenia Risk Loci to Synaptic Function." Schizophrenia Bulletin 42 (6): 1334–42.

Sekar, Aswin, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, et al. 2016. "Schizophrenia Risk from Complex Variation of Complement Component 4." Nature 530 (7589): 177–83.

Shi, Yichen, Peter Kirwan, James Smith, Hugh P. C. Robinson, and Frederick J. Livesey. 2012. "Human Cerebral Cortex Development from Pluripotent Stem Cells to Functional Excitatory Synapses." Nature Neuroscience 15 (3): 477–86, S1.

Skene, Nathan G., Julien Bryois, Trygve E. Bakken, Gerome Breen, James J. Crowley, Helena Gaspar, Paola Giusti-Rodriguez, et al. 2017. "Genetic Identification Of Brain Cell Types Underlying Schizophrenia." https://doi.org/10.1101/145466.

Stein, Jason L., Luis de la Torre-Ubieta, Yuan Tian, Neelroop N. Parikshak, Israel A. Hernández, Maria C. Marchetto, Dylan K. Baker, et al. 2014. "A Quantitative Framework to Evaluate Modeling of Cortical Development by Neural Stem Cells." Neuron 83 (1): 69–86.

Stiles, Joan, and Terry L. Jernigan. 2010. "The Basics of Brain Development." Neuropsychology Review. https://doi.org/10.1007/s11065-010-9148-4.

Sullivan, Patrick F., Arpana Agrawal, Cynthia Bulik, Ole A. Andreassen, Anders Borglum, Gerome Breen, Sven Cichon, et al. 2017. "Psychiatric Genomics: An Update and an Agenda." https://doi.org/10.1101/115600.

Tai, Yu Chuan, and Terence P. Speed. 2006a. "A Multivariate Empirical Bayes Statistic for Replicated Microarray Time Course Data." Annals of Statistics 34 (5): 2387–2412.

"A Multivariate Empirical Bayes Statistic for Replicated Microarray Time Course Data." Annals of Statistics 34 (5): 2387–2412.

Tanapat, Patima. 2013. "Neuronal Cell Markers." Materials and Methods 3. https://doi.org/10.13070/mm.en.3.196.

Tang, Xin, Li Zhou, Alecia M. Wagner, Maria C. N. Marchetto, Alysson R. Muotri, Fred H. Gage, and Gong Chen. 2013. "Astroglial Cells Regulate the Developmental Timeline of Human Neurons Differentiated from Induced Pluripotent Stem Cells." Stem Cell Research 11 (2): 743–57.

The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. 2017. "Meta-Analysis of GWAS of over 16,000 Individuals with Autism Spectrum Disorder Highlights a Novel Locus at 10q24.32 and a Significant Overlap with Schizophrenia." Molecular Autism 8 (1): 21.

Wood AR Esko T Yang J Vedantam S Pers TH Gustafsson S Chu AY Estrada K Luan J Kutalik Z Amin N Buchkovich ML Croteau-Chonka DC Day FR Duan Y Fall T Fehrmann R Ferreira T Jackson AU Karjalainen J Lo KS Locke AE Mägi R Mihailov E Por, Frayling T. M. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." Nature Genetics 46 (11): 1173–86.

Zhang, Y., K. Chen, S. A. Sloan, M. L. Bennett, A. R. Scholze, S. O'Keeffe, H. P. Phatnani, et al. 2014. "An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex." Journal of Neuroscience 34 (36): 11929–47.

Chapter 3 - Supplemental Materials

Full supplemental information can be found here:
https://doi.org/10.1016/j.biopsych.2018.08.019

S1. Supplementary Material and Methods

S1.1 Approval for stem cell research

The University of California, Los Angeles Embryonic Stem Cell Research Oversight (ESCRO) committee approved this work. Their policy is based on the recommendations of the National Bioethics Advisory Commission, the National Academies of Science-Institute of Medicine guidelines, and standards created by the California Institute for Regenerative Medicine.

S1.2 Human neural stem cell line

WA09(H9)-derived hNSC is a commercially available and commonly used neural stem line with standardized and well-documented neuronal differentiation protocols(1–3). These cells originate from a donated human embryo (F), produced by *in vitro* fertilization for clinical purposes, that was cultured to a blastocyst after which an ESC line was established(4, 5). This cell line is of European ancestry(6) and has a normal karyotype. It was in addition successfully tested for stem cell characteristics and approved by NIH for stem cell research(7). WA09 ESCs were differentiated to NSCs by the vendor and obtained by us as neural progenitors. Tissue culture plates were coated with CELLstart CTS™ (Thermo Fisher Scientific) diluted (1:50) in DPBS with Ca^{2+} and Mg^{2+} and hNSCs cells expanded in KnockOut™ DMEM/F-12 Basal Medium (Gibco) with 2% StemPro® Neural Supplement (Gibco). 2mM GlutaMax™-I Supplement (Gibco), FGF Basic and EGF Recombinant proteins (Gibco, both at 20 ng/ml), and 1x Pen Strep (Thermo Fisher Scientific). Cells were plated at $1.0x10^5$ cells per 3.8 cm^2, dissociated with preheated StemPro Accutase (Gibco) and subsequently passaged at ~90% confluency. This cell line tested negatively for mycoplasma contamination both at the vendor
and in our lab.

S1.3 Experimental design and RNA extraction

Cells all originated from the same batch of hNSCs differentiated from the WA09 hESC line. We specifically chose to perform our experiments across an isogenic background to minimize variation and maximize statistical power to identify transcriptomic signatures across differentiation. Each sample was cultured in a separate well and represents an independent differentiation process, which makes for semi-technical replicates. After RNA extraction, samples were quantified using the Quant-iT™ RiboGreen® RNA Assay Kit (Thermo Fisher Scientific). RNA integrity was assessed through RIN scores using the Agilent 2100 Bioanalyzer (mean +/- sd = 9.26 +/- 0.63). Transcriptome data was collected at the UCLA Neuroscience Genomics Core using Illumina's HumanHT-12 v4 Expression BeadChip Kit, which is a cost-effective platform.

S1.4 Data preprocessing and quality control

We select for probes present in at least 1 sample at detection p-value of <0.01. Probes were in addition filtered for quality by "perfect" or "good" annotation using the illuminaHumanv4.

db package (v1.26) in R. Network adjacency by Euclidean distance and standardized connectivity (Z.K) were calculated on filtered probes values using the WGCNA package to detect outliers, defined as having Z.K. < -2(8, 9). All samples survived this exclusion threshold. As RNA samples were randomized across gene expression arrays, batch has no explanatory value on days of differentiation (R^2=0.0, p=1.0, see also Figure S1).

S1.5 in vitro cellular identity

An RNA-sequencing (RNA-Seq) transcriptome database of major classes of cell types present in the cerebral cortex was used to assess cell type-specific gene expression across neuronal differentiation. Briefly, gene expression data of purified populations of neurons, astrocytes, oligodendrocyte precursor cells (OPC), newly formed oligodendrocytes (NFO), myelinating oligodendrocytes (MO), microglia, and endothelial cells from mouse cerebral cortex was downloaded from the database(10). Fold changes in gene expression values, using fragments per kilobase of exon per million fragments mapped (FPKM), for each gene in each cell type were compared to the mean expression level across the other six cell types. To enrich for cell type-specific genes, we selected the top genes sorted by fold change, with a minimal fold change of 2 and FPKM < 5 in the other brain cell types.

S1.6 Transition mapping to a spatiotemporal atlas of early human brain development

To investigate global transcriptomic matching between *in vitro* gene expression profiles and *in vivo* gene expression profiles of neocortical brain regions, we applied transition mapping (TMAP)(11). This method uses a spatiotemporal transcriptome atlas of the human brain(12) and laminar expression data dissected via Laser Capture Microdissection from fetal human brain as *in vivo* input(13). Both data sets contain brain samples from multiple individuals. TMAP only includes neocortical regions in the analyses. The method performs serial differential gene expression (DGE) analysis between any developmental stages or cortical laminae in the *in vivo* datasets and DGE analysis between two *in vitro* time points of choice. Both DGE lists are sorted on –log10(p-value) and multiplied by the sign of the beta coefficient from the DGE analysis. TMAP subsequently implements the Rank Rank Hypergeometric Overlap (RRHO) test to determine overlap between the *in vitro* and *in vivo* DGE ranked lists and produces RRHO Difference maps that visualizes the extent of overlap(14). The TMAP and RRHO analyses are implemented in the online CoNTExT bioinformatic pipeline (https://context.semel.ucla.edu). Analyses were run for *in vitro* time points day-0 vs day-30, day-0 vs day-5, day-5 vs day-15, and day-15 vs day-30 across both temporal and spatial dimensions of human cortical development.

S1.7 Time-series differential gene expression analysis

Two multivariate empirical Bayes models are used to identify differentially expressed genes across *in vitro* neuronal differentiation. The first method exploits the correlation structure among time points and replicates to identify non-constant genes and applies moderation by borrowing the information across genes into the analyses to reduce type-I and type-II errors due to poorly estimated variance-covariance matrices(15). This method is implemented in the Timecourse package (v 1.42) in R. We used the mb.long() function to calculate the one-sample

T2 statistic that ranks genes based on their log10 probability to have differential expression over time. The second method, Bayesian Estimation of Temporal Regulation (BETR), is an extension of the first approach and uses a flexible random-effect model that allows for correlations between the magnitude of differential expression at different time points(16). This method explicitly models the joint distribution of the samples across time points and calculates the probability of a gene being differentially expressed using Bayes rule. BETR is implemented in the betr package (v 1.26) in R. These two methods complement each other as the first approach has increased sensitivity for transient expression differences while BETR has increased sensitivity to detect genes with non-constant expression that is small but sustained over multiple consecutive time points(16). To maximize our power to detect differentially expressed genes across time points and replicates, we applied both methods to rank genes by their probability of having non-constant gene expression across *in vitro* neuronal differentiation.

S1.8 Fuzzy c-means cluster analysis

Fuzzy c-means clustering is a soft clustering approach that allows probes to obtain fuzzy memberships to all clusters, minimizes the effect of noise in the data, and avoids erroneous detection of clusters generated by random gene expression patterns. Fuzzy c-means clustering is performed in Euclidian space on standardized gene expression values. This ensures that genes with similar changes in expression cluster together. Membership values represent cluster affiliations and highlight the extent of similarity in expression between genes. To calculate cluster membership values, we first have to estimate a fuzzifier, which determines the level of cluster fuzziness, and the optimal cluster number to use. These two parameters were empirically estimated from the data (fuzzifier = 1.55, number of clusters = 8) as previously described using the Mfuzz package in R(17, 18) (Figure S20). We used these two optimal estimates and subsequently calculated cluster membership with the fclusList() and membership() function in the Mfuzz package. Because these functions only take gene expression values of a single-replicate time series as input, we randomly sampled 100 single-replicate time series from our data and calculated cluster membership values using standardized gene expression values for each independent time series (Figure S21). We then proceeded to calculate average cluster membership for each probe for each cluster across our 100 independently sampled time series (Figure S22). These average cluster membership values were then used for all downstream analyses.

S1.9 Functional annotation of clusters

The Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6. 8) was used for functional annotation of each cluster(19). We restricted our analysis to probes with high membership, i.e. cluster membership > 0.5, to identify most informative functional annotations (Table S1). At a membership value of > 0.5, there is no overlap in genes between clusters (Figure S23). With this setting, 4,318 genes were assigned to a cluster with an average cluster size of 540 genes with the smallest and largest cluster having 221 and 891 genes, respectively. DAVID was run using unique Ensembl IDs and the following databases: UP_KEYWORDS, UP_SEQ_FEATURE, GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT, BIOCARTA, KEGG_PATHWAY, INTERPRO, UCSC_TFBS. Genes significantly detected during differentiation (n = 12,414) were

set as background to determine gene overrepresentation in clusters. The functional annotation clustering tool was applied at default settings to group gene list with overlapping gene IDs. Cluster annotations were called significant if the enrichment > 1.0 and at least 1 gene list in the annotation cluster survived Bonferroni correction ($P < 0.05$).

S1.10 Intolerance of loss-of-function variation across clusters

The probability of being loss-of-function (LoF) intolerant (pLI) was used to infer functional gene constraint across clusters. pLI measures were downloaded (April 2017) for 18,225 genes from the ExAC Browser (http://exac.broadinstitute.org/downloads). The statistical framework underlying the pLI metric is described by others in more detail elsewhere(20). The Wilcoxon Rank-Sum test was used to test if cluster constraint was statistically different between groups.

S1.11 Generation of annotation files for MAGMA and sLDSC

To integrate GWAS data with *in vitro* transcriptomic signatures, annotation files mapping Ensembl gene IDs or chromosomal position to *in vitro* gene parameters were created. Illumina probe IDs were mapped to Ensembl gene IDs using NCBI build 37.3, duplicate IDs removed, and gene boundaries extended symmetrically by 10kb to include regulatory regions. Continuous gene parameters used are the T2-statistic and cluster membership values, which are first collapsed per gene ID using the mean values across probes and subsequently log-transformed and rank-transformed, respectively. For binary annotation files, we assigned genes or chromosomal positions a 1 or 0 for being above or below a specified threshold, respectively. These files were then used as input to Multi-marker Analysis of GenoMic Annotation (MAGMA) and stratified LD score regression (sLDSR) to integrate *in vitro* signatures with GWAS data and study the cumulative impact across risk loci.

S1.12 GWAS summary statistics used

GWAS summary statistics were checked and reformatted using the munge_sumstats.py program within the ldsc software, which removes low quality and ambiguous variants(21). SNPs in the MHC region (hg19 - chr6: 28477797 – 33448354) were filtered out due to extensive linkage disequilibrium (LD) between markers in this region. The APOE locus (hg19 – chr19: 44,409,039–46,412,650) was removed from analysis of AD to minimize the effect of variants with large effect sizes in downstream regression analyses. For MDD, we included GWAS results from the China Oxford and VCU Experimental Research on Genetic Epidemiology (CONVERGE) consortium(22) and 23andMe Inc., a personal genetics company(23). The latter uses a proxy of self-reported depression as a phenotype. We did not include the MDD GWAS of the PGC(24) in our analyses as it has a strong genetic correlation with the self-reported depression GWAS (rg=0.72)(23) but a lower h^2 z-score. The 1000 Genomes Project Phase 3 release (1KG) was used as reference panel to model ancestry-matched LD(25). We used 503 individuals of European ancestry and 301 individuals of East Asian ancestry in analyses of GWAS data derived from target population of Europeans and Han Chinese, respectively.

S1.13 MAGMA gene-set analysis

MAGMA (v1.06)(26) was used to perform gene-set analyses of GWAS data, which uses a multiple regression framework to associate a continuous or binary gene variable to GWAS gene level p-values. For each GWAS trait, we generated gene-level p-values by computing the mean SNP association using the default gene model ('snp-wise=mean'). SNPs were mapped to genes using Ensembl gene IDs and NCBI build 37.3 with +/- 10kb extensions of gene boundaries using the --annotate flag. We only included SNP with minor allele frequency (MAF) > 5% and dropped synonymous or duplicate SNPs after the first entry ('synonym-dup=drop-dup'). For each annotation, we then regressed gene-level GWAS test statistics on the corresponding gene annotation variable using the '--gene-covar' function while adjusting for gene size, SNP density, and LD-induced correlations ('--model correct=all'), which is estimated from an ancestry-matched 1KG reference panel. In all analyses, we included only genes for which we had both the gene variable and GWAS gene level test statistic available. Testing only for a positive association, i.e. enrichment of GWAS signal, we report one-sided p-values along with the corresponding regression coefficient.

S1.14 Stratified LD Score Regression - generating annotation files and LD scores

For sLDSR, we used a recent extension to the method that partitions h^2 by continuous-valued annotations(27). This extension relies on the assumption that if a continuous annotation is associated to increased h^2, LD to SNPs with larger values of this annotation will increase the χ^2 statistic of a SNP more than LD to a SNP with smaller values. We first generated sLDSR annotation files and computed LD scores for each continuous-valued annotation. We mapped gene log(T^2-statistic) and standardized cluster memberships to SNPs in 1KG reference panel BIM files. To increase the number of SNPs in our analyses, we extended gene boundaries with 100kb on each end, similar to here(28). SNPs that intersected with a gene were annotated with the corresponding gene variable, while SNPs that did not map to genes were annotated with zero. For each annotation, we then estimated partitioned LD scores using using the ldsc.py --l2 function with MAF > 5% and a 1 centimorgan (cm) window. As recommended, only HapMap3 SNPs IDs, with the MHC region removed, were written and used in the final regression model. In case of binary gene annotations, a 1 (in the annotation) and 0 (not in annotation) coding was used. In a similar fashion, we computed LD scores for all 53 annotations in the baseline model (see Supplementary Methods for details). We in addition generated weight files that contain non-partitioned LD scores using only SNPs that will be included in the final regression model. These are LD scores computed from the HapMap3 SNPs with the MHC region removed. Frequency files were generated with the -- freq flag in PLINK 1.9(28–30).

We next generated baseline annotation files using BED files of 52 functional annotations, which were downloaded from the LDSC web portal. Genomic interval coordinates in each BED file were intersected with SNPs present in 1KG reference panel BIM files. If a SNP intersected with an interval in a BED file it was annotated as 1 for that particular annotation. If a SNP did not intersect, it was annotated as 0. In addition to 52 annotations, we also added a recommended base annotation that coded a 1 for every SNP. These 53 annotations makeup the baseline model. With the generated sLDSR annotation files and 1KG reference panels we estimated LD scores

for each annotation using the ldsc.py --l2 function with MAF > 5% and a 1 centimorgan (cm) window. As recommended, only HapMap3 SNPs, with the MHC region removed, were written and used in downstream analyses. As a sanity check, we correlated our estimated CEU baseline LD scores to the baseline LD scores that can be downloaded from the LDSC web portal and found a high concordance. For example, the mean Pearson correlation between computed LD scores across baseline annotations on chromosome 22 is 0.99 (n=53, sd=0.002). Thus, we proceeded and used the baseline model in our analyses as it has been shown to provide more accurate mean estimates of enrichment. The baseline model and the details of each annotation are described elsewhere(28, 31).

S1.15 Stratified LD Score Regression – main model

We ran sLDSR (ldsc.py --h^2) for each annotation of interest while accounting for the full baseline model and an extra annotation of all genes detected in our *in vitro* model (n = 12,414). That is, for each annotation we ran the following model;

1. Full baseline model with 53 annotations.
2. Annotation of all genes detected during *in vitro* neuronal differentiation.
3. Annotation of interest (e.g. cluster membership).

If an annotation of interest (3) is associated with increased h^2, LD to SNPs with large values of that annotation will increase the c^2 statistic of a SNP more than LD to SNPs with smaller values. To determine if this effect is significant and specific to this annotation, it estimates the contribution of that annotation to the per-SNP h^2 while accounting for the baseline and the all genes detected annotation (1 + 2). As we only test for a positive association, we report the contribution to the per-SNP h^2 (τ) and the associated one-sided p-value, which is calculated using standard errors that are obtained via a block jackknife procedure.

S1.16 CORTECON replication

The CORTECON human cortical differentiation dataset, described in detail somewhere else(32), was used for replication analysis. Briefly, H9 human ESCs were induced to neural progenitors over a course of 12 days. On day 13, neural progenitor induction medium was changed to a culture medium supporting cortical differentiation and cells further differentiated up to 77 days. RNA was extracted at day nine time points (0, 7, 12, 19, 26, 33, 49, 63, and 77 days) in duplicates and paired-end RNA-seq data collected. FASTQ files were mapped to the human genome (hg19) and raw count data made available through the CORTECON repository (http://cortecon.neuralsci.org/). Raw RNA-seq data for 44,562 transcripts and 24 samples were downloaded from the repository and analyzed in R. Transcripts with at least ten reads in more than two samples (n=19,008) were called significantly detected and retained for further analyses. The count data was subsequently read into a DEseqDataSet object, then normalized by size factors and transformed by stabilizing the variance over the mean (vst() function) using the DESeq2 package in R(33). Entrez Gene IDs were mapped to Ensembl Gene IDs and duplicates collapsed by their average transformed expression value. Transformed values were used as input for the PCA analysis, visualization of marker gene expression, and fuzzy c-means clustering as described

for the discovery analyses. To identify differentially expressed genes, the nbinomLRT() was applied to test for significance of change in deviance between a full model (gene xp ~ 1 + Time) and a reduced model (gene xp ~ 1) for each gene. The model first estimates size factors and dispersion and then uses the likelihood ratio test (LRT) to assess if the increased likelihood of the data using the full model is more than expected given the reduced model. For the purposes of replication and comparison between datasets, only genes significantly detected in both datasets were used in the analysis (n=11,290).

S1.17 Statistical Analysis

Statistical analyses were performed with R (https://www.r-project.org) or an otherwise specified algorithm. Significance with MAGMA and sLDSR was determined at a one-sided α level of 0.05. If applicable, Bonferroni correction for multiple comparison is applied and denoted in figures and tables.

Supplementary Figures and Results

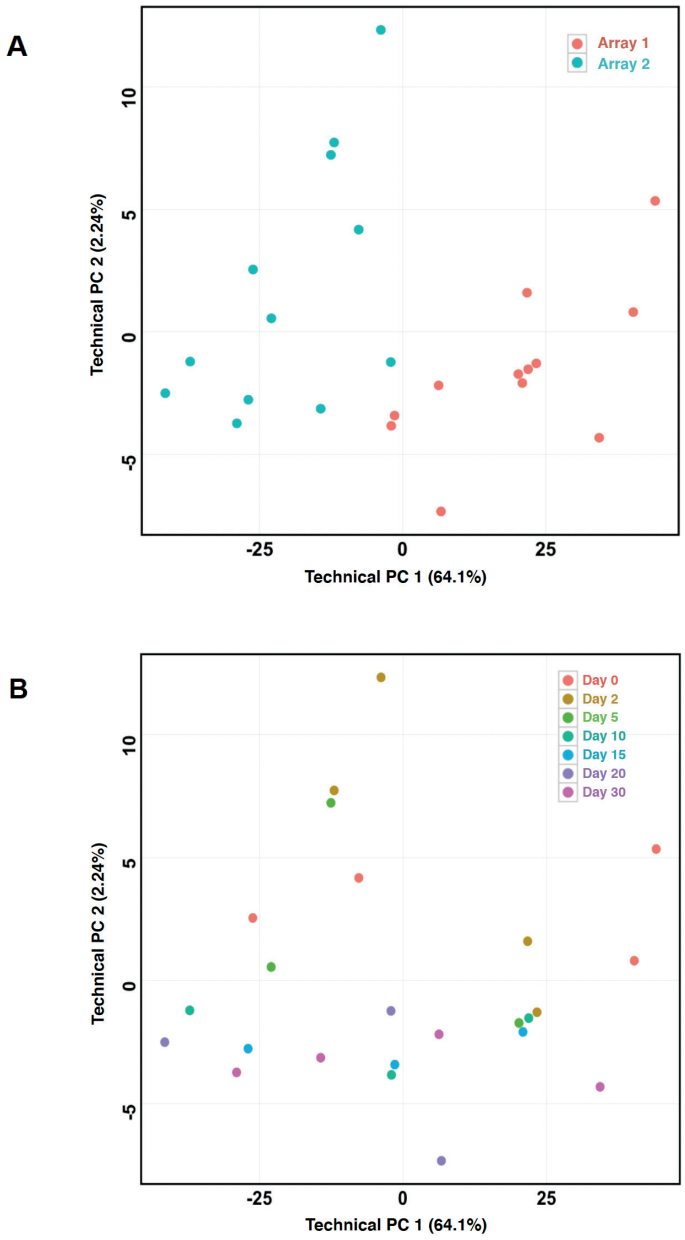


Figure S1. Results of PCA on control probes. The human HT-12 v4 beadchip contains 887 control probes that capture technical variation. Plotted above are PC1 and PC2 with variance explained in parentheses. Dots in the graphs represent samples and are color-coded by (A) array and (B) time. PC1 explains the majority of the information of the control probes but has no correlation with time in culture.

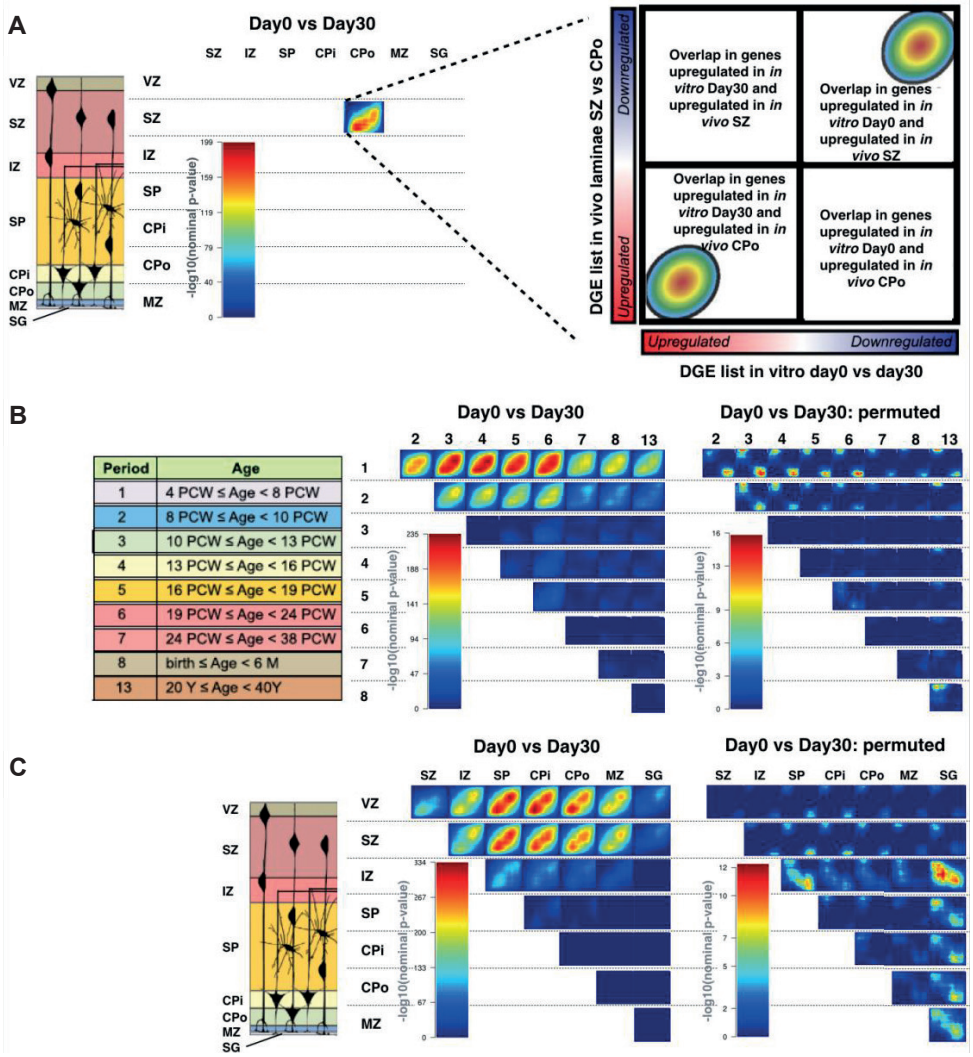


Figure S2. Gene expression overlap between in vitro neuronal differentiation and in vivo human cortical development. *CoNTExT was used to apply transition mapping and generate Rank Rank Hypergeometric Overlap difference maps. (A) Shows a toy example of how to interpret difference maps of overlap between in vivo time points and in vivo laminae. in vitro day-0 vs day-30 differential gene expression (DGE) profile was mapped to serial DGE profiles of (B) human brain developmental stages and (C) laminae of the human cerebral cortex. Difference maps show the amount of matching between in vitro and in vivo DGE profiles. Maps are colored by –log10(p-value) denoted by each corresponding color bar. On the right of (B) and (C), results are also shown for analyses with permuted in vitro sample labels. Abbreviations and numbering above maps correspond to schematic representations on the left (adopted from Stein et al., 2014) of different developmental stages and laminae. VZ = ventricular zone, SZ = subventricular zone, IZ = intermediate zone, SP=subplate zone, CPi= inner cortical plate, CPo = outer cortical plate, MZ = marginal zone, PCW = post conception weeks, M = months, Y = years, Period = developmental stage.*

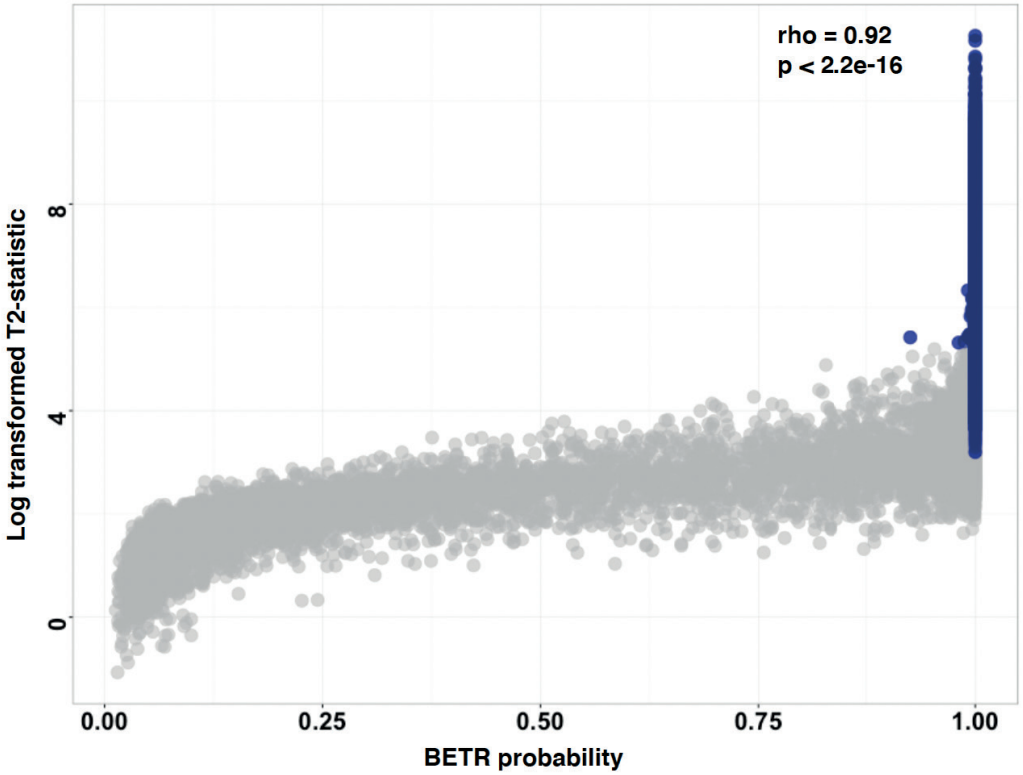


Figure S3. A scatterplot showing the concordance between two methods that identify non- constant genes over time. *The x-axis shows the probability from BETR. The y-axis shows the log transformed T^2 statistic from the second method. Each dot represents a probe. Blue color indicates the union of probes that are confidently called as having non-constant expression over time (n=7,734). The Spearman correlation between the ranks is shown in the top right corner.*

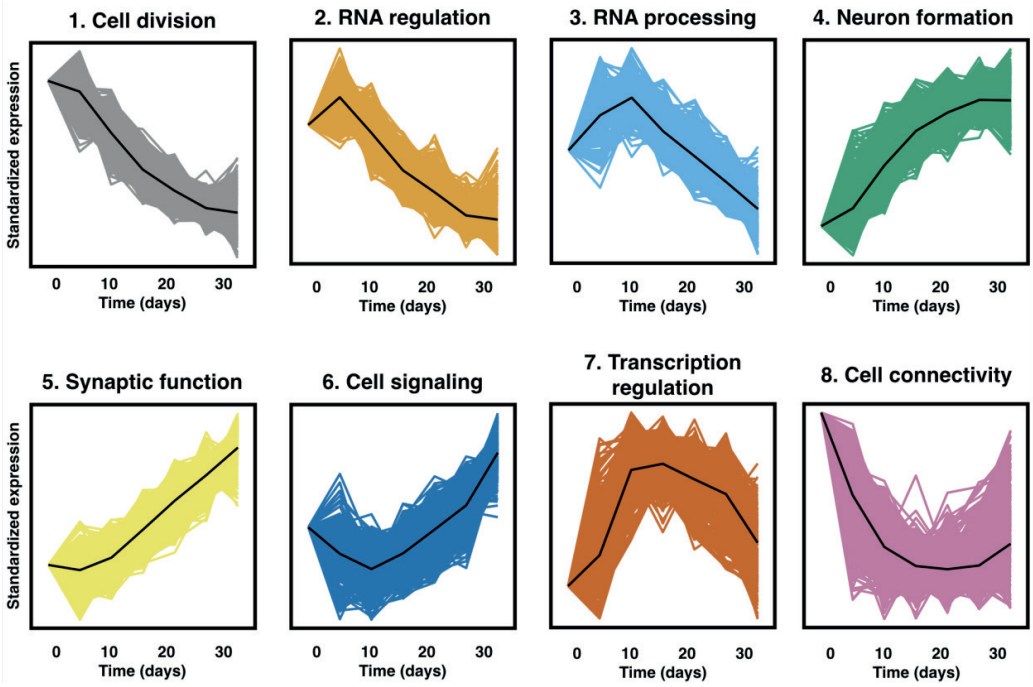


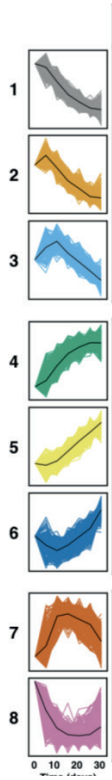
Figure S4. Experimentally-derived longitudinal gene clusters. *An enlarged representation of gene expression patterns of high confidence gene members for each cluster (see also figure 3). The x-axis denotes the time across differentiation and the y-axis gene expression values standardized to day-0. The black line highlights the average expression patterns of each cluster.*

S2. Supplementary Results

S2.1 Upregulated genes are more likely to be intolerant to loss-of-function functional variation

Recent work has shown that intolerance to loss-of-function (LoF) functional variation (i.e. constraint) in genes and gene sets can highlight core biological processes and likelihood of disease pathogenicity(20, 34). High constraint genes have been implicated in neurodevelopmental disorders, such as autism spectrum disorder (ASD) and intellectual disability(34), and are in addition more likely to be adjacent to GWAS signal than the average gene(20). We therefore investigated constraint across clusters and extracted probabilities of LoF intolerance (pLI) from the ExAC database(20). The median pLI across all 18,225 genes extracted from the browser is 0.027. Differentially expressed genes (n=5,545, median pLI=0.285) have increased average gene constraint compared to non-differentially expressed genes (n=6,839, median pLI=0.085). This difference between the groups is significant ($W=2.09x10^{7}$, $P < 2.2x10^{-16}$). Genes that are upregulated during differentiation primarily drive the increase in constraint. More specifically, genes in clusters that are affiliated to *neuronal maturation* (median pLI = 0.55, n=633) and s*ynaptic function* (median pLI = 0.52, n=616) show a significant increase in pLI while genes affiliated to *cell division* (median pLI = 0.067, n=543), RNA binding (median pLI = 0.046, n=285), and *extracellular matrix* (median pLI = 0.104, n=490) show a significant decrease in pLI relative to differentially expressed genes (see Figure S5 for test statistics). This shows that genes that are upregulated during neuronal differentiation have a lower tolerance to functional disruption than the average gene expressed, which makes these genes interesting to study in the context of disease.

Supplementary Figure S5-8



	Genes (n)	Functional annotations	Enrichment	Cluster label	LoF constraint (median pLI)	LoF constraint (p-value)
1	737	Mitotic cell cycle and division Chromosome and kinetochore DNA repair	23.90 17.41 15.67	Cell division	0.067 ↓	8.54E-11
2	457	RNA binding and ribonucleoproteins RNA processing and ribosome biogenesis RNA splicing	15.77 10.05 6.18	RNA regulation	0.046 ↓	7.64E-06
3	221	RNA processing and binding ncRNA processing *Regulation of cell cycle*	4.25 3.40 *2.13*	RNA processing	0.166	0.48
4	891	Neuronal outgrowth Synaptic formation Neuron differentiation	6.29 4.60 4.23	Neuron formation	0.551 ↑	7.54E-08
5	779	Neuronal projections Synaptic signaling Synapse	16.85 13.56 10.16	Synaptic function	0.524 ↑	1.02E-06
6	253	Membrane-bound vesicles Glycoprotein signaling Transmembrane region	4.17 3.95 2.72	Cell signaling	0.180	0.20
7	396	Regulation of transcription Transcription factor activity Zinc finger and ion binding	4.21 3.02 2.97	Transcription regulation	0.207	0.35
8	584	Extracellular region Cell junctions and adhesion Regulation of cell motility	8.20 7.78 4.85	Cell connectivity	0.104 ↓	3.31E-05

Figure S5. Genes upregulated during neuronal differentiation are intolerant for loss-of-function genetic variation. Cluster annotations shown with average gene constraint and its association with gene cluster membership shown across clusters.

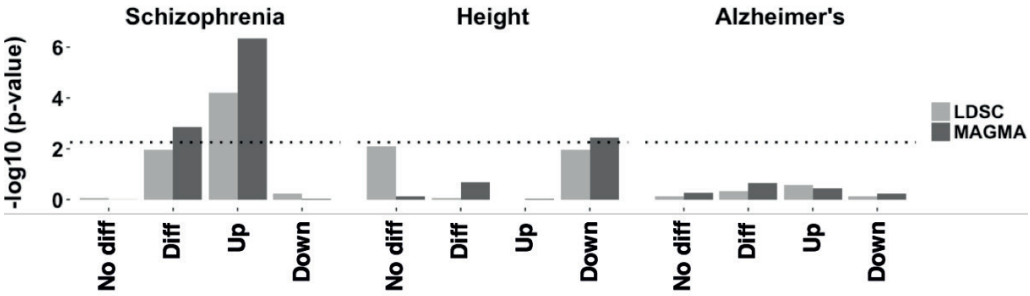


Figure S6. Height and Alzheimer's disease show no h^2 enrichment in up-regulated genes. *A more detailed investigation of the enrichment of h^2 of SCZ, height, and Alzheimer's disease across differentially expressed genes. The y-axis denotes the –log10 P-value of the enrichment. No diff = genes that are not differentially expressed; Diff = log (T^2-statistic) as shown in Table 1; Up = genes up-regulated during differentiation; Down = genes down-regulated during differentiation. The dotted line represents the threshold for P = 0.0056 (n=9 tests).*

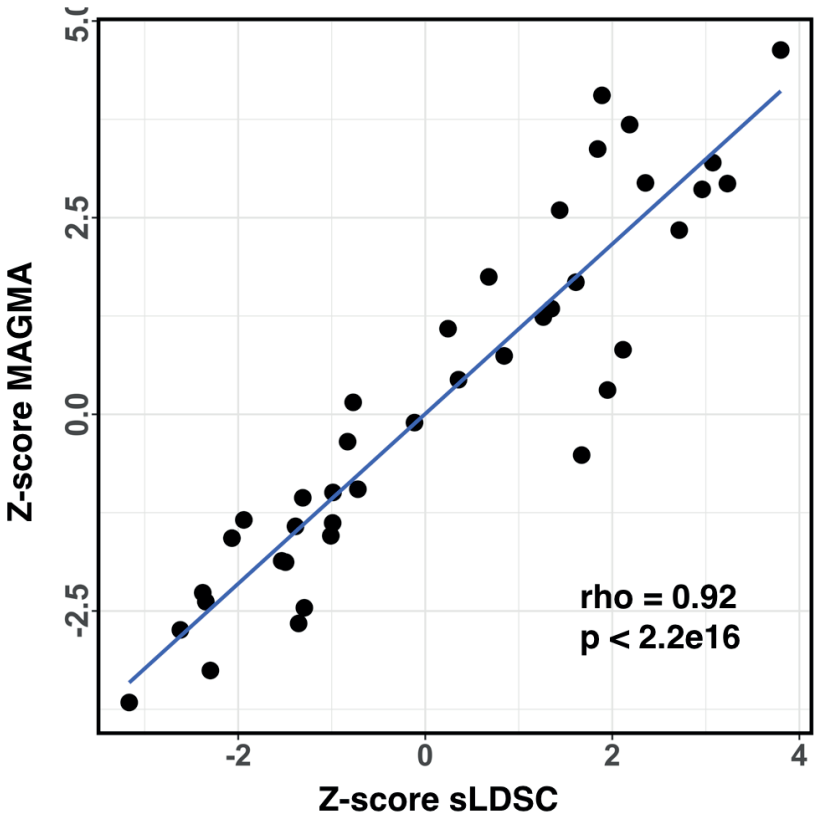


Figure S7. MAGMA and sLDSC show strong concordance in results. *Each dot represents the results of phenotype-cluster combination for both MAGMA (y-axis) and sLDSC (x-axis) (n=40). The regression line is shown in blue with the Spearman correlation between the ranks in the bottom right corner.*

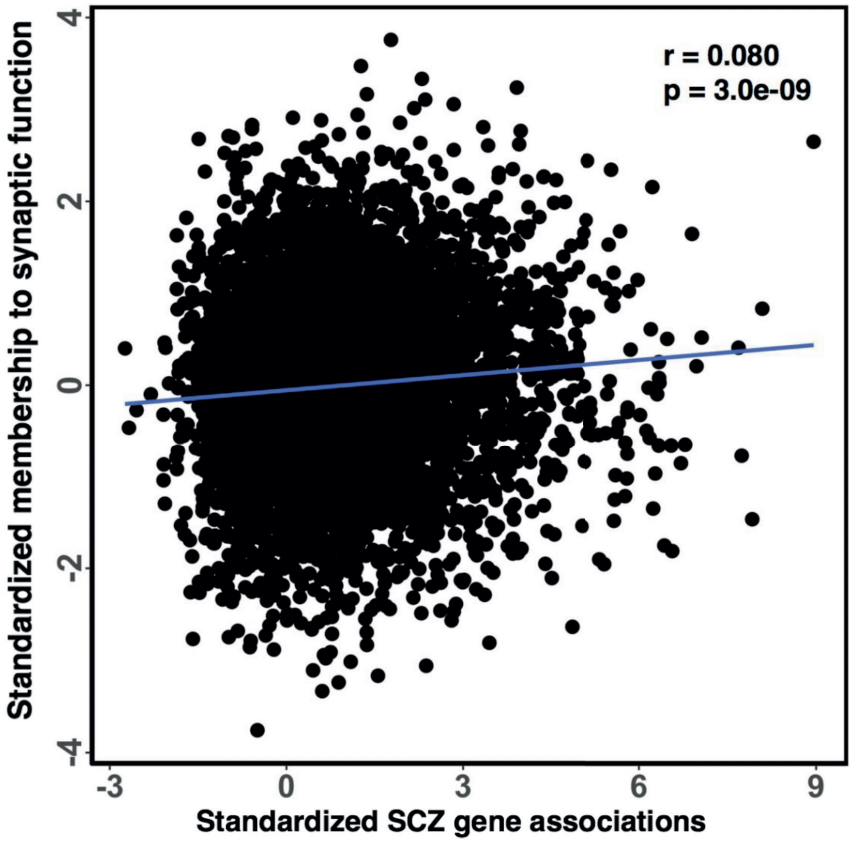


Figure S8. A plot showing the association between SCZ gene-level association statistics and synaptic cluster gene membership. *Standardized membership values to the synaptic function cluster and standardized gene level association statistics are shown on the y-axis and x-axis, respectively. The regression line is shown in blue with Pearson correlation test statistics denoted in the top right corner. The plotted association is not yet corrected for gene size, SNP density nor LD.*

S2.2. Cluster enrichments of schizophrenia and height are inversely correlated

For height, we found effects in opposite direction of psychiatric traits in the downregulated gene clusters. We find an inverse correlation between enrichments of SCZ and height across eight gene clusters (rho=-0.86, P=0.011, n=9, see also Figure S9), despite the absence of any evidence of a genetic correlation across the whole-genome (rg=-0.002, p=0.95)(35). Our findings however do suggest a genetic correlation. Indeed, large-scale epidemiological studies have, for example, reported an inverse relationship between adult height and SCZ(36, 37). A population-based cohort study of >1 million Swedish men describes a 15% reduction in SCZ risk for tall subjects compared to short subjects(37). It has therefore been suggested that height and SCZ are likely to have overlapping genetic causes that can be both discordant and concordant(38). Our results are in line with this hypothesis and suggest that discordant and concordant effects aggregate on pathway levels that are dependent on time and place during development (Figure S10). While future work is needed to further explore the genetic relation between SCZ and height, these observations illustrate the added value of individual longitudinal gene clusters and highlight a complex genetic relationship between these two phenotypes.

Supplementary Figure S9-11

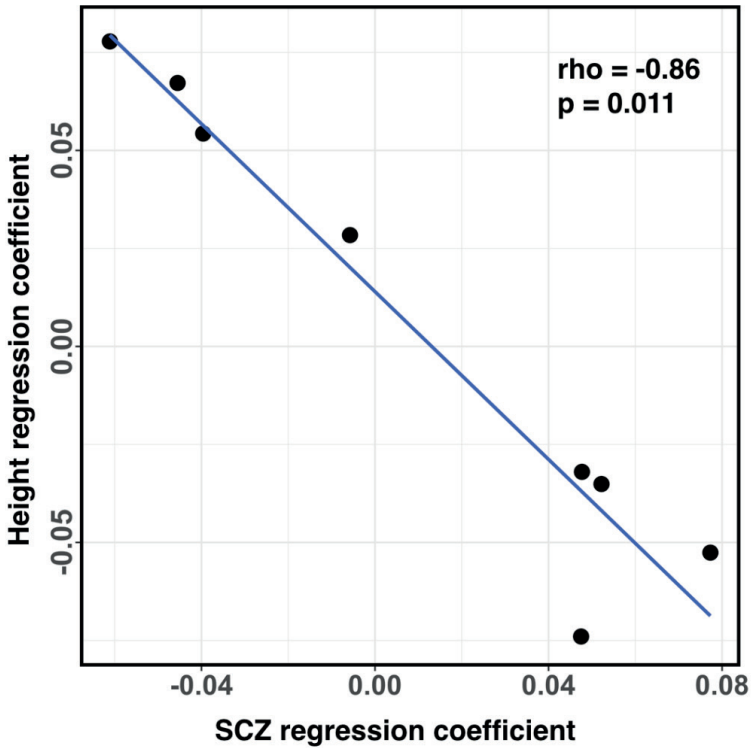


Figure S9. Schizophrenia and height show an inversely correlated pattern of enrichment results. *Shown are MAGMA results with each dot representing the regression coefficients of enrichment for schizophrenia and height on the x-axis and y-axis, respectively. The Spearman correlation between the ranks of both methods is shown in the top right corner along with the corresponding significance level.*

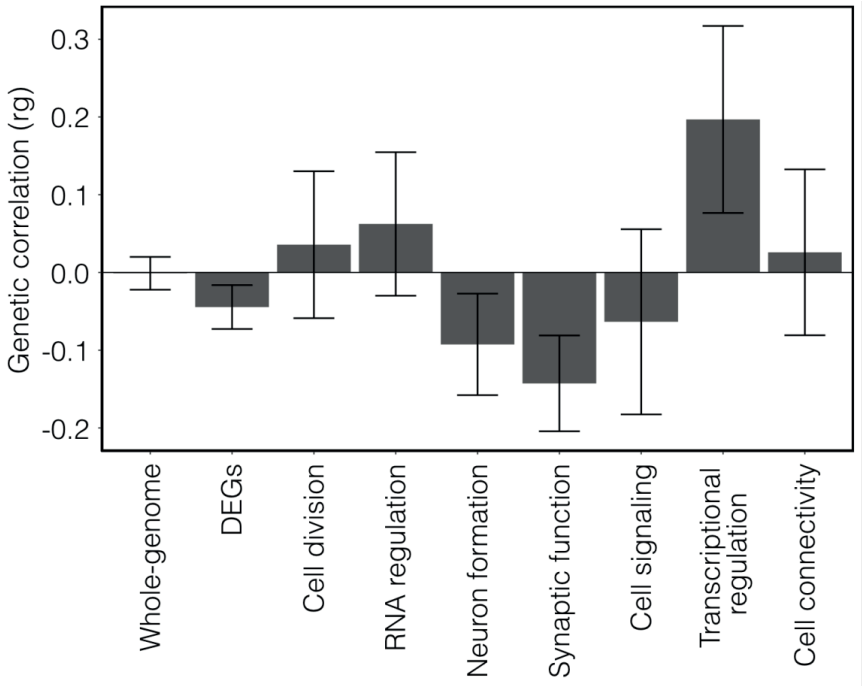


Figure S10. The genetic correlation between schizophrenia and height varies across cluster while absent across the whole genome. *Genetic correlations were determined using cross-trait LD score regression and SNPs with MAF > 5%. Stratified correlations were computed using only a subset of SNPs that overlap with genomic coordinates of the highest gene members of that cluster (membership > 0.5). For one cluster (RNA processing), the subset of SNPs was too few to compute a genetic correlation. For differentially expressed genes (DEGs), the correlation was computed on SNPS overlapping the union of DEGs (n=5,818). Error bars represent the standard error.*

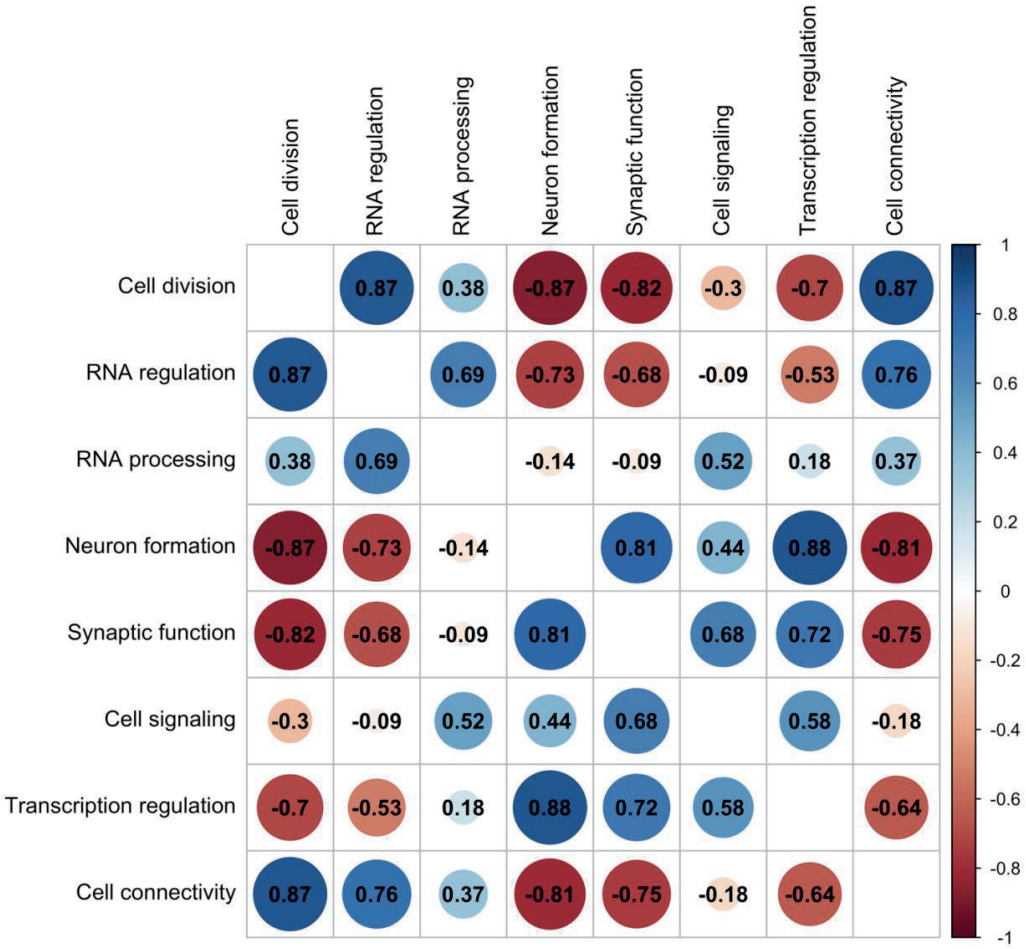


Figure S11. The correlation structure across clusters. *A matrix with spearman's correlations calculated between gene membership values across clusters. The rho is denoted in each cell and the strength of the correlation color coded according to the bar on the right.*

S2.3. Replication analysis in the CORTECON RNA-Seq dataset (Figure S12-17)

S2.3.1 Comparison of dataset characteristics and differences

We first compared the main characteristics and experimental variables between both dataset (Table 1). Similar to the discovery dataset, the CORTECON project used the WA-09 stem cell line(32). However, for the discovery data we obtained human *neuronal stem cells* (hNSCs) and differentiated progenitors to a neuronal fate up to 30 days, whereas the authors of the CORTECON study started at the *embryonic stem cell state* and performed neural progenitor induction themselves. After induction, they subsequently differentiated to a dorsal telencephalic fate up to 77 days. Furthermore, differentiation protocols, RNA isolation procedures, time points and number of replicates collected, and gene expression platforms used are different. We in addition implemented a data processing and analysis pipeline that is different from our discovery analysis to accommodate the use of RNA-seq data. While we emphasize that these datasets are very different, we did embark on assessing the reproducibility of marker gene expression levels, global transcriptomic signatures, and SCZ GWAS enrichment in the CORTECON dataset.

	Discovery dataset	**CORTECON dataset**
Cell line	WA-09 hNSC	WA-09 hESC
Days of differentiation	30 days	77 days
Time points	0, 2, 5, 10, 15, 20, 30	0, 7, 12, 19, 26, 33, 49, 63, 77
ESC culture:		
coating	Manufacturer's protocol	Grown on MEFs
medium	Manufacturer's protocol	HES-medium
Neural progenitor culture:		
induction	Manufacturer's protocol	Day1-12; cyclopamine, N2
coating	CellStart	Matrigel
medium	StemPro NSC SFM +FGF/EGF	KSR-medium + LDN193189/SB431542
Neuronal culture:		
induction	Day 1	Day 13
coating	Poly-D-Lysine/Laminin	Matrigel
medium	Neurobasal B27	Neurobasal B27 + N2 + FGF2
RNA extraction:	Qiagen's Allprep kit	Qiagen's RNAprotect + RNeasy kit
Transcriptomic technology	Array	RNA-sequencing

Table Note. Dataset characteristics and experimental settings between discovery and replication dataset. While both dataset use the WA-09 cell line, they are different in many ways. Further details on the CORTECON dataset can be found in van de Leemput & Boles et al., 2014.

S2.3.2 Transcriptome correlation analysis highlights time-specific similarity between datasets

To gain insight in the degree of comparability between dataset and assess transcriptomic similarity between differentiation trajectories, we performed a correlation analysis between all pair of samples across datasets. Using overlapping genes significantly expressed and in the top quartile of variable genes across time points in each dataset (n=856), we computed the Spearman rank correlation across all pairs and visualized these in the heatmap below (Figure S12). We find that as differentiation progresses the transcriptomic similarity between datasets follows the differentiation trajectory over time. For example, day-30 of differentiation in the discovery dataset shows a negative correlation with the earliest time points in the CORTECON dataset. The observed correlation gradually shifts to a positive correlation as neurons develop over time, with day-77 in the replication having the largest positive correlation with day-30 in the discovery dataset (mean rho = 0.31). This demonstrates that genes that change most over time and that are expressed in both dataset share significant similarity in their expression levels across the trajectory of differentiation between experiments. To align datasets for subsequent analyses, we aimed to select a time window in the CORTECON dataset that best matches the 30-day *in vitro* differentiation trajectory of the discovery experiment. As we started our differentiation at the neural progenitor stage (day-0), we chose day-12 in CORTECON as the first time point as this is the end of neural progenitor induction (day-13 is start of neuronal induction). Day-0 of discovery and day-12 of CORTECON also display a strong correlation based on gene expression levels (mean rho = 0.51). We selected two endpoints of differentiation in CORTECON; day-49 (37 days of neuronal differentiation) and day-77 (65 days of neuronal differentiation), with the first mapping most closely to the number of days of differentiation in our discovery dataset (i.e. 30 days).

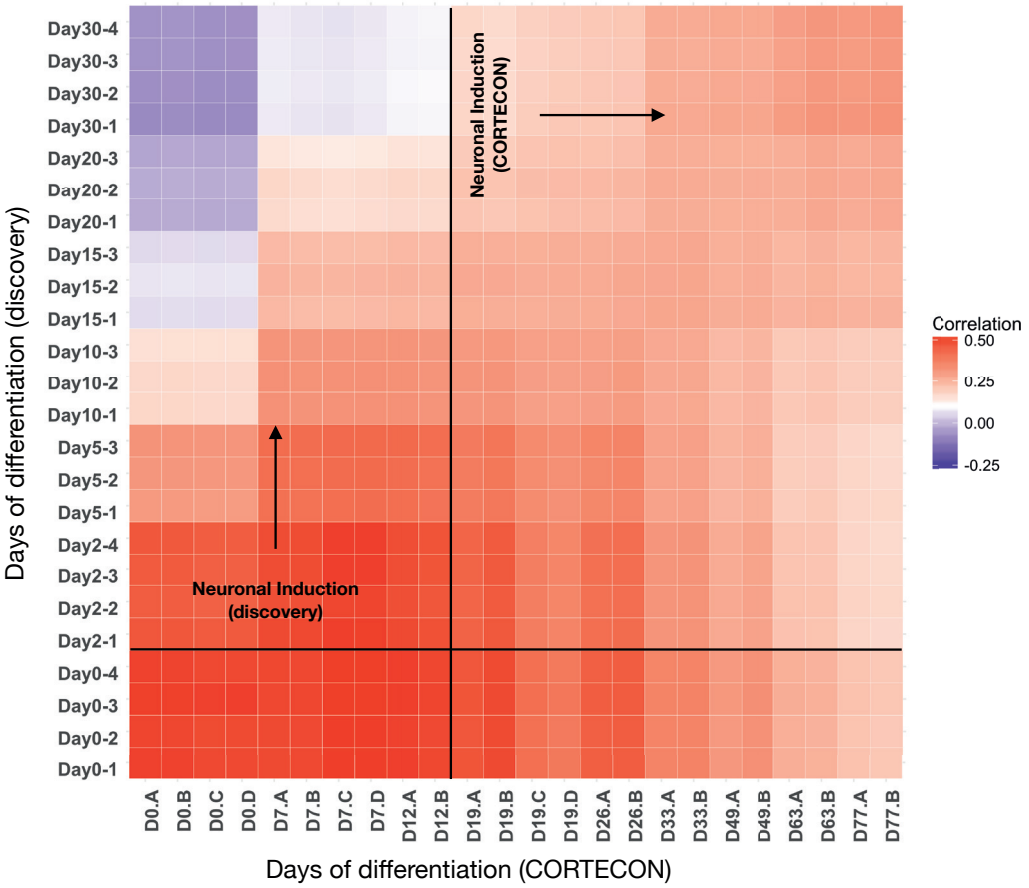


Figure S12. Transcriptome correlation analysis reveals similarity of differentiation trajectories between datasets. To gain insight in the degree of comparability between dataset and assess transcriptomic similarity between differentiation trajectories, we performed a correlation analysis between all pair of samples across datasets. The x-axis shows samples of the CORTECON dataset ordered by days of differentiation and replicate. The y-axis shows the discovery dataset. The Spearman rank correlation is shown across a color gradient. The black lines and arrows indicate the starting point of neuronal induction for each dataset.

S2.3.3 Principal component analysis highlights the differentiation trajectory

We investigate global transcriptomic signatures across differentiation using principle component analysis (PCA). PCA is performed using all time points and replicates (n=24) and variance stabilized, normalized counts of 16,791 protein-coding genes that passed QC. We find that the first PCs explain 67% of the data and capture the differentiation process quite accurately (Figure S13; left plot). Zooming in on day12-77, the start of neuronal induction until the end of differentiation, we find that some samples do display an irregular pattern (middle plot, highlighted by black circle) and deviate from the expected smooth gradual transition on the PC-axis as differentiation progresses. This may be due to a batch effect or sample mix-up/ contamination at some point during sample processing. Based on the alignment of the samples on the axis of the first PC using the full dataset (left plot), we excluded two replicates of day 19 (samples day19-C/D). We do believe there is still a batch effect in the data but are not able to track this back accurately and thus account for it. For further analysis within day12-77, this leaves seven time points with two replicates each (n=14; right plot).

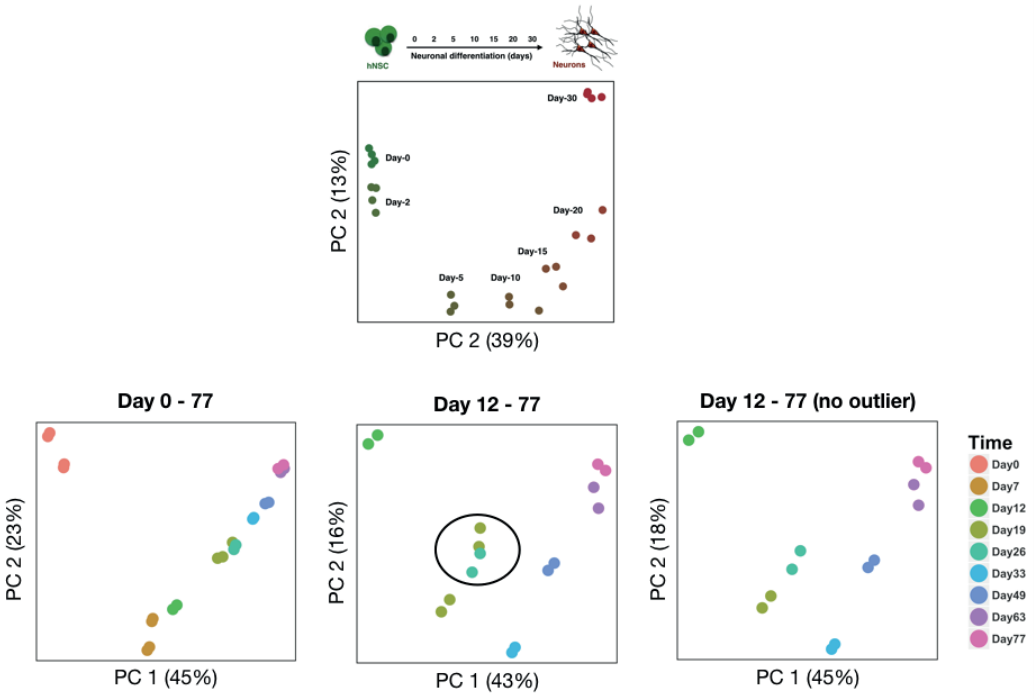


Figure S13. in vitro gene expression profiles capture the differentiation trajectory in the CORTECON dataset. Shown are the PCA plot of the discovery dataset (top) with three PCA plots of three subsets of the CORTECON dataset (bottom) with principle compontent (PC) 1 and PC2 plotted with variance explained in parentheses on the axis labels.

S2.3.4 Traditional gene markers show consistent patterns of expression across datasets

To confirm that the differentiation trajectories are in line with the implemented protocols, we examined traditional neural stem cell and neuronal marker genes and their expression patterns over days of differentiation. We identified three stem cell markers (MKI67, Nestin, and SOX2) and four neuronal markers (BDNF, DCX, TUBB3, and MAP2). These markers are significantly detected in both dataset and commonly reported to evaluate the differentiation trajectory(39, 40). Using normalized expression levels; we standardized each gene's expression values with the first time point at zero and a standard deviation of one. We then plotted each marker gene's standardized expression values over time for both the discovery and replication dataset. We find that the classical neural stem cell markers Nestin and SOX2 and MKI67, a marker of proliferation, are downregulated over time as cells differentiate away from their progenitor state towards a neuronal lineage (Figure S14). This observation is consistent between both datasets and in line with what has been reported.

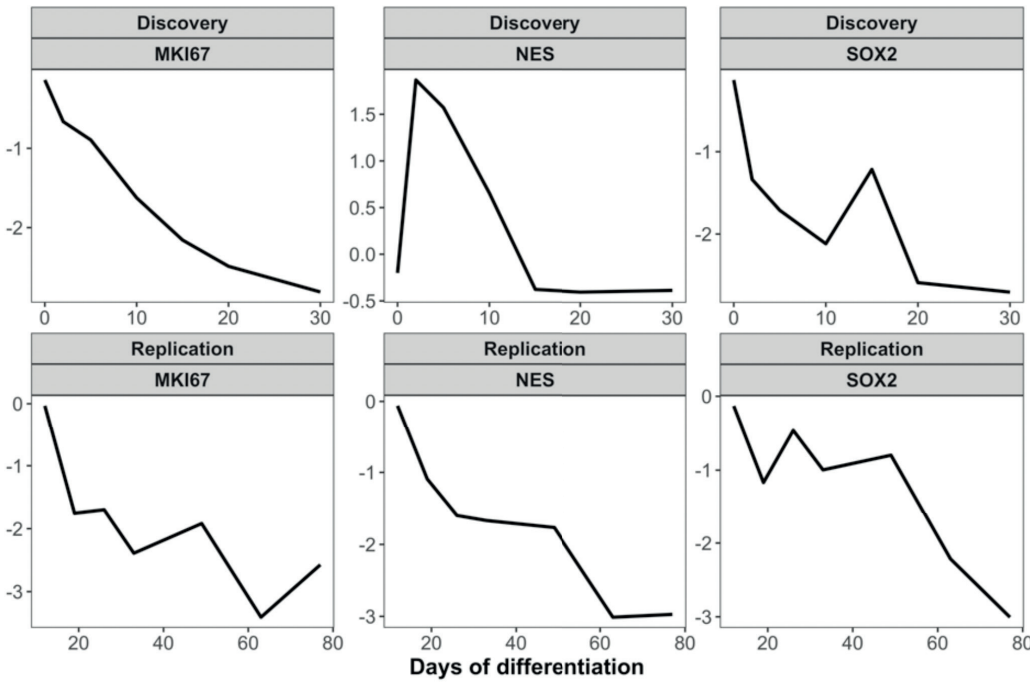


Figure S14. Stem cell and proliferation marker genes are downregulated over time in both datasets. *Shown are standardized gene expression values of three traditional neural stem cell marker genes plotted over time. The x-axis shows the days of differentiation. Neuronal induction was induced at day-1 in the discovery and day-13 in the replication. MKI67 - proliferation marker protein Ki-67; NES - Nestin; SOX2 - sex determining region Y-box 2.*

Next, we examined four neuronal markers and find a consistent pattern between datasets of upregulated expression as neuronal differentiation progresses for BDNF and DCX,

which are early markers of neuronal differentiation and developing neurons (Figure S15)(41). In the CORTECON dataset, we find an indecisive pattern of expression for TUBB3 and MAP2 as differentiation progresses over time that does not align with the more gradual expression pattern in the discovery dataset. This may suggest that the effect of their neuronal induction protocol may be more variable or heterogenous and affecting only a subset of genes involved in specific (and more mature) domains of neuronal functioning. It may also reflect the irregular pattern we observed in the PCA analysis (Figure S13), although it remains speculative. Taken together, we do believe that both datasets largely show the same results, which is that neural stem cell markers are downregulated and early neuronal markers upregulated as differentiation progresses. Based on the expression of TUBB3 and MAP2 in both datasets, we conclude that these neuronal cultures remain however immature, which limits the developmental range of the model. Having said that, an important next experiment would be to implement strategies to improve the maturity of the culture, for example with co-culturing of astrocytes.

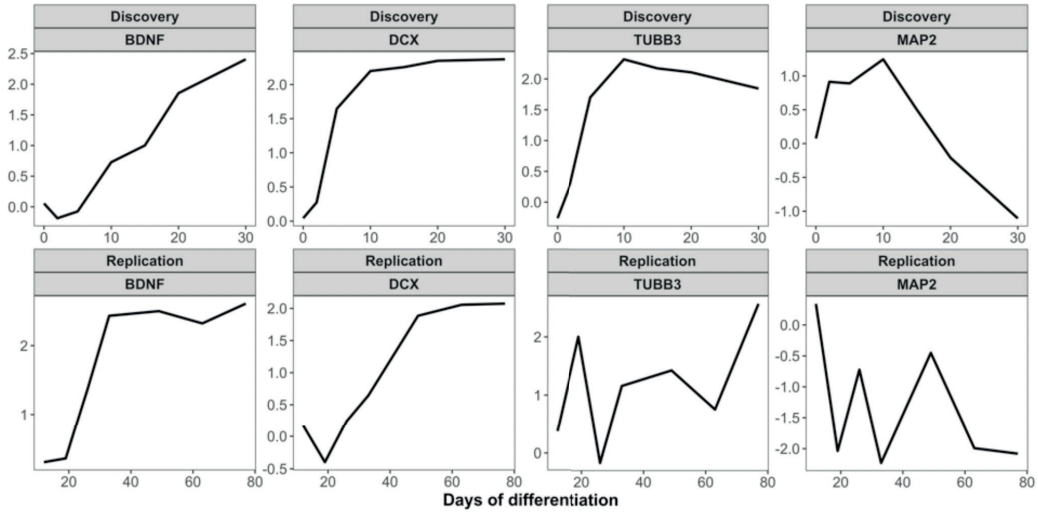


Figure S15. Early neuronal marker genes are upregulated over time in both datasets. *Shown are standardized gene expression values of three traditional neural stem cell marker genes plotted over time. The x-axis shows the days of differentiation. Neuronal induction was induced at day-1 in the discovery and day-13 in the replication. BDNF - brain-derived neurotrophic factor; DCX - doublecortin; TUBB3 - tubulin beta-3 chain; MAP2 – microtubule associated protein 2.*

S2.3.5 Differentially expressed genes show significant overlap between datasets

To identify genes differentially expressed across differentiation, we applied a likelihood ratio test (LRT) using the DESeq2 gene differential expression pipeline(33). We choose to deviate from the bioinformatics pipeline in our discovery analysis as the methods to identify differentially expressed genes are specifically designed for microarray gene expression time series data. As RNA-seq data consists of counts that follow a negative binomial distribution, these methods are

not suitable for the CORTECON time series dataset. The DESeq model, for each gene, implements the LRT to compare (1) a full model [gene counts ~ 1 + time] against (2) a reduced model [gene counts ~ 1]. The LRT determines if the increased likelihood of the data using the full model with the time variable (model 1) is more than expected if the time component is truly zero (model 2). While the DESeq method is a widely used pipeline to analyze RNA-seq data, we do acknowledge that it relies on a linear regression model and is therefore less powered to identify genes with nonlinear expression across differentiation. Methods to analyze time series RNA-seq data are however scarce and this framework will allow us to identify genes differentially expressed in the CORTECON dataset and compare with the genes identified in our discovery analysis. We analyzed two differentiation windows in the CORTECON dataset; day 12-49 (n=10, 37-days of neuronal differentiation) and day 12-77 (n=14, 65-days of neuronal differentiation). For day 12-49, we identify 7,379 genes out of 16,791 genes to be differentially expressed over time at FDR < 5% of which 4,905 are also significantly detected in the discovery dataset. Of the 4,905 genes, we find that 2,739 (56%) are also identified as differentially expressed in our discovery analysis, which is a highly significant overlap that is unlikely to happen by chance (P=9.8e-26, OR = 1.51, Table 2). When we analyze the 65 days of neuronal differentiation window (day 12-77), we find 9,869 genes differentially expressed at FDR < 5% of which 6,267 are also detected in our discovery dataset. Of these 6,267 genes, 3,314 (53%) are also differentially expressed in our discovery analysis, which is a significant overlap (P=4.03e-10, OR=1.28, Table 2) but to a lesser extend as the genes found within the 37-day differentiation window.

CORTECON Day 12-49

Discovery	OR=1.51; P=9.8e-26	Differentially expressed	Not differentially expressed	Totals
	Differentially expressed	2,739	2,516	5,255
	Not differentially expressed	2,166	3,001	5,167
	Totals	4,905	5,517	10,422

CORTECON Day 12-77

Discovery	OR=1.28; P=4.0e-10	Differentially expressed	Not differentially expressed	Totals
	Differentially expressed	3,314	1,941	6,267
	Not differentially expressed	2,953	2,214	4,155
	Totals	5,255	5,167	10,422

Table note: Time series gene differential analysis results in the CORTECON dataset significantly overlaps with discovery. *Shown are contingency tables that were used as input for the Fisher's exact test to assess the overlap in differentially expressed genes between discovery and replication dataset. Shown are the results for day 12-49 (top) and day 12-77 (bottom) with the odds ratio and the significance level of the test denoted in the upper left corner.*

These results demonstrate that genes that are important for *in vitro* neuronal differentiation significantly overlap between datasets. Moreover, we find that the 37-day differentiation window in the CORTECON dataset maps more closely to the transcriptome of the 30-day differentiation window of our discovery dataset than the 65-day differentiation window. This is most likely driven by different transcriptomic signatures that are present at later differentiation time points and highlights the importance of temporal alignment and specificity when comparing neuronal differentiation transcriptome datasets. We next performed time series cluster analysis to group differentially expressed genes into longitudinal gene clusters that are active over time.

S2.3.6 CORTECON gene clusters show significant overlap with discovery dataset

We applied our time series clustering pipeline and identified six and seven longitudinal gene clusters for 37 days of differentiation and 65 days of differentiation, respectively. After computing mean cluster membership values for each differentially expressed gene for each cluster, we identified genes with high degree of membership to a cluster (membership > 0.5) and used these to identify functional annotation enrichments via DAVID (v6.0). Visualized below (Figure S16) are the results, including the overlap between high membership cluster genes between 37-days and 65-days and the top functional annotation associated with each cluster.

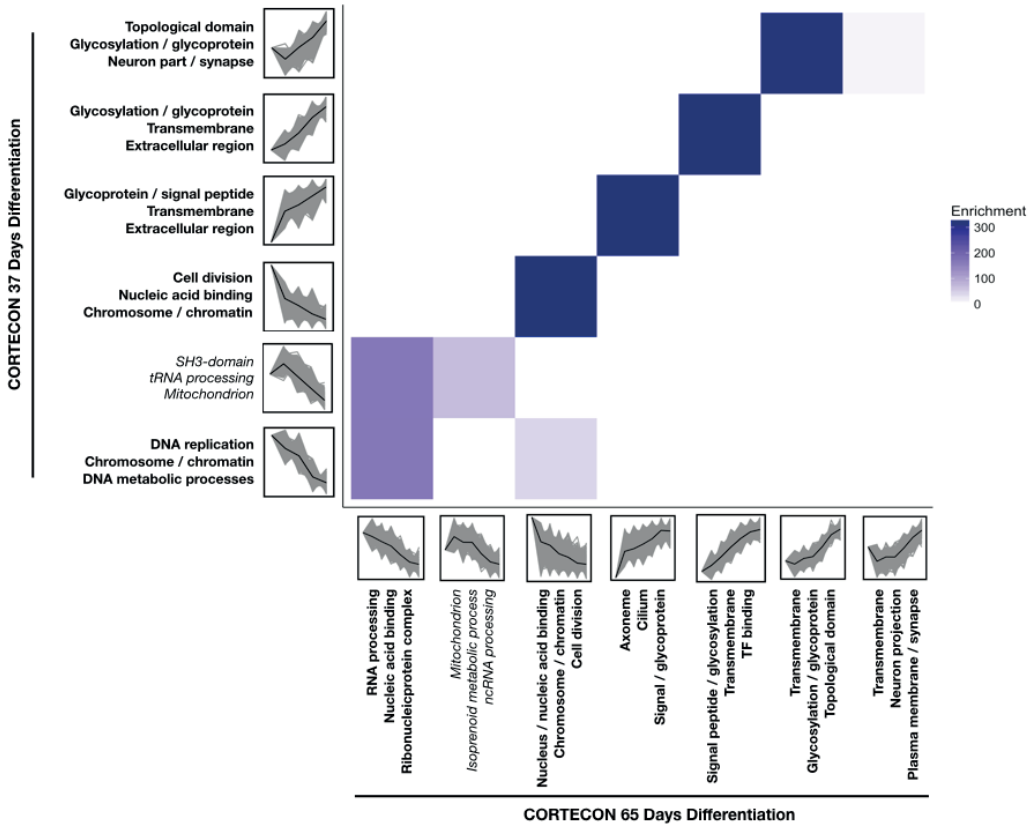


Figure S16. Time series gene expression cluster analysis in the CORTECON dataset. *Time series gene expression cluster analysis in the CORTECON dataset. Shown are the gene clusters identified for both 37-days of differentiation (y-axis) and 65-days of differentiation (x-axis) with the top three functional annotations denoted. Gene overlap between clusters is shown, with strength of overlap visualized by the outcome of the hypergeometric overlap test (-log10 p-value).*

Similarly to the discovery dataset, we find gene clusters that are downregulated over time and enriched for cell division, RNA processing, and chromosome organization functional annotation. Likewise, we find clusters that are upregulated over time that are associated with neuronal functioning. Across 65-days of differentiation, we find an additional cluster that is upregulated during late differentiation that is not present in the 37-day differentiation window. Note that we do not detect any clusters with nonlinear expression patters, which we did in the discovery analysis. This is a limitation of the DESeq2 regression framework. Interestingly, we observe similar top functional annotations across upregulated clusters for both day-37 and day-65. Glycosylation and glycoproteins and transmembrane annotations particularly stand out. Glycosylation refers to a post-translational modification of protein and lipids by monosaccharides or oligosaccharide chains and plays an important role during vertebrate development, including regulation of the nervous system. While we did find a similar cluster in our discovery analysis, it

was not as abundantly distributed across all upregulated clusters. This may be due to specific differences in differentiation protocols which additional studies may provide biological insights into. To compare the CORTECON gene clusters with the discovery gene clusters, we investigated the overlap between clusters identified during 37 days of differentiation in CORTECON and 30 days of differentiation in the discovery analysis. We find that gene clusters that are downregulated over time show significant overlap between datasets (Figure S17). Similarly, upregulated gene clusters show a significant degree of overlap. As noted before, nonlinear gene clusters in the discovery dataset show weak or no overlap with clusters identified in CORTECON. This likely does not reflect true differences between datasets but is an outcome of differential bioinformatic pipelines used to identify differentially expressed genes. Overall these findings do highlight that transcriptomic signatures that are important for *in vitro* neuronal differentiation are largely shared across datasets and that there are also finer differences within gene clusters and the functional annotations that they associate with.

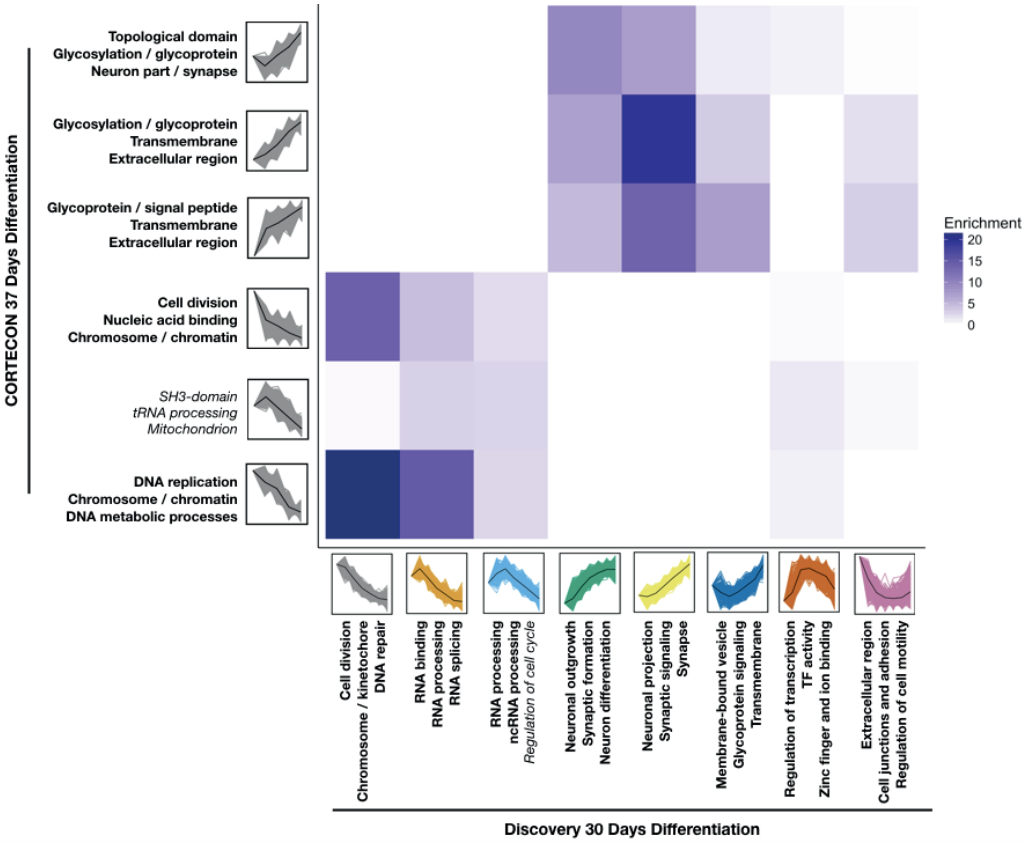


Figure S17. Gene clusters identified in CORTECON overlap significantly with cluster of the discovery analysis. Shown are the gene clusters identified for both 37-days of differentiation (y-axis) and our 30-days of differentiation in the discovery (x-axis) with the top three functional annotations denoted. Gene overlap between clusters is shown, with strength of overlap visualized by the outcome of the hypergeometric overlap test (-log10 p-value).

S2.3.7 SCZ polygenic risk is significantly associated with CORTECON differently expressed genes that are upregulated over time

In our discovery analysis, we found that SCZ polygenic risk is significantly associated with differentially expressed genes that are upregulated over time during differentiation (beta=0.12, P=9.2E-07). This association is specifically driven by a longitudinal gene cluster that is enriched for synaptic function annotations (yellow cluster in Figure S17). We performed a similar two-step analysis in the CORTECON dataset, where we first associated SCZ polygenic risk with differentially expressed genes and subsequently followed up on identified gene clusters. Using MAGMA, we replicate our finding that genes differentially expressed during 37 days of differentiation in CORTECON are significantly associated with SCZ risk (beta=0.047, P=0.007, Table 3). This is, similar to the discovery dataset, driven by genes that are upregulated over time (P=0.008) and not downregulated (P=0.74). We find no association of SCZ risk with differentially expressed genes across 65 days of differentiation in CORTECON.

SCZ GWAS Enrichment	Beta (SE)	Beta_std	P-value
Discovery 30 days			
Differentially expressed	0.022 (0.007)	0.094	0.001
Upregulated	0.120 (0.025)	0.069	9.2E-07
Downregulated	-0.054 (0.025)	-0.030	0.984
CORTECON 37 days			
Differentially expressed	0.047 (0.019)	0.033	0.007
Upregulated	0.056 (0.023)	0.030	0.008
Downregulated	0.014 (0.023)	0.008	0.740
CORTECON 65 days			
Differentially expressed	0.004 (0.020)	0.003	0.847
Upregulated	-0.002 (0.022)	-0.001	0.929
Downregulated	0.006 (0.021)	0.004	0.756

Table note: SCZ polygenic risk is enriched in genes differentially expressed across 37 days of in vitro neuronal differentiation. *Shown are results of MAGMA using differentially expressed genes as gene-set for our discovery analysis, CORTECON 37 days of differentiation (day12-49), and CORTECON 65 days of differentiation (day12-77). Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given being in the gene-set as compared to out of the gene-set.*

We examined the association within 37 days of differentiation further by analyzing the six gene clusters we identified but find no evidence of the association with SCZ risk to be distributed across a specific gene clusters.

SCZ GWAS Enrichment	Beta (SE)	Beta_std	P-value
CORTECON 37 days			
Cluster 1	0.008 (0.016)	0.009	0.311
Cluster 2	0.003 (0.015)	0.003	0.432
Cluster 3	0.004 (0.015)	0.005	0.393
Cluster 4	-0.021 (0.015)	-0.024	0.918
Cluster 5	-0.019 (0.015)	-0.022	0.894
Cluster 6	-0.020 (0.015)	-0.022	0.900

Table note: SCZ polygenic risk association in CORTECON is not distributed to a specific gene cluster. Shown are results of MAGMA with rank-transformed gene membership values of each cluster as predictor of SCZ GWAS gene-level z-scores. Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given 1SD in standardized membership value.

To investigate whether similar genes are driving the association with SCZ risk between our discovery analysis and the differentially expressed genes across 37 days of the CORTECON dataset, we adjusted our analysis in the CORTECON dataset for the synaptic gene cluster (n= 779 genes) of the discovery analysis. We find that the strength of the association between SCZ risk and day-37 differentially expressed genes that are upregulated decreases when we account for synaptic genes from the discovery analysis (beta=0.044, P=0.031, see table below). This suggests that, in part, similar genes underlie the association between SCZ GWAS risk and transcriptomic signatures across *in vitro* neuronal differentiation between both datasets.

SCZ GWAS Enrichment	Beta (SE)	Beta_std	P-value
CORTECON 37 days			
Differentially expressed	0.047 (0.019)	0.033	0.007
Upregulated	0.056 (0.023)	0.030	0.008
Downregulated	0.014 (0.023)	0.008	0.740
CORTECON 37 days - Synaptic genes adjusted			
Differentially expressed	0.044 (0.019)	0.031	0.015
Upregulated	0.044 (0.023)	0.024	0.036
Downregulated	0.022 (0.022)	0.012	0.843

Table note: SCZ polygenic risk association in CORTECON is in part driven by synaptic genes identified in discovery. Shown are results of MAGMA with differentially expressed genes as gene-set with and without correction for the synaptic genes identified in the discovery analysis (n=779). Beta = regression coefficient, SE = standard error, Beta_std = change in Z-value given being in the gene-set as compared to out of the gene-set.

Supplementary Figure S18-23

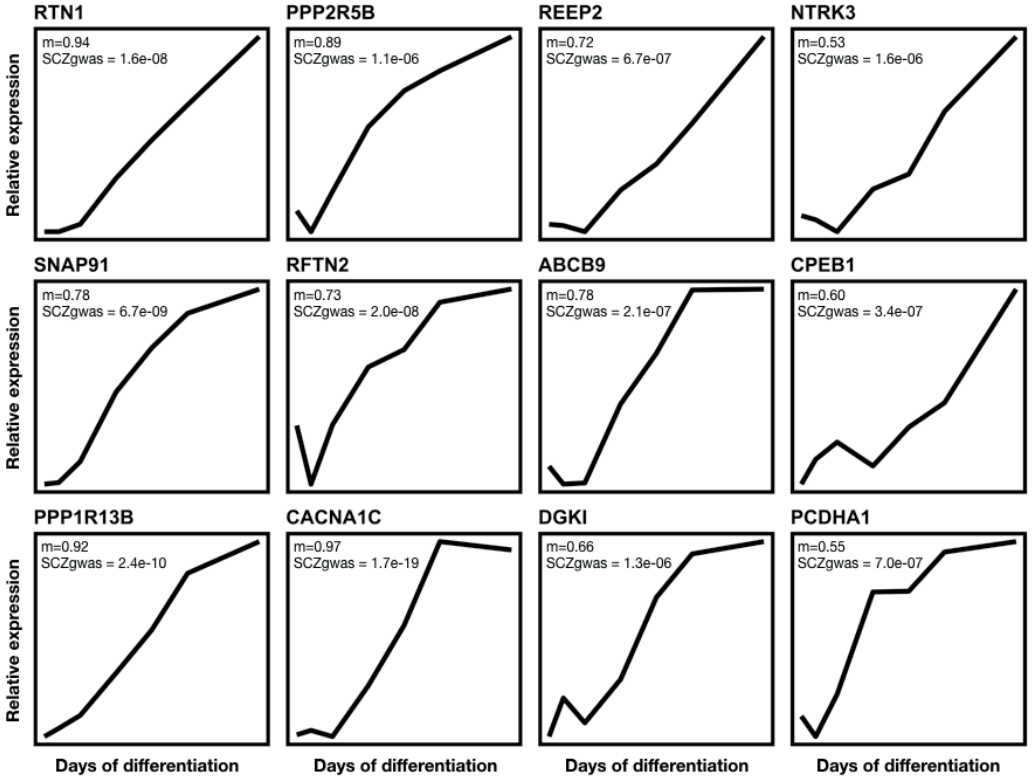


Figure S18. High membership synaptic genes significantly associated with SCZ risk. Shown are relative expression patterns over time in the discovery datasets of genes with high membership (m > 0.5) to the synaptic function gene cluster. These genes (n=12) are also identified to be significantly differentially expressed in the CORTECON dataset (FDR < 5%) and have a SCZ GWAS gene-level p-value < 2.5e-06 (Bonferroni correction). For each gene, we show the gene symbol alongside its synaptic gene cluster membership value (m) and SCZ GWAS gene p-value.

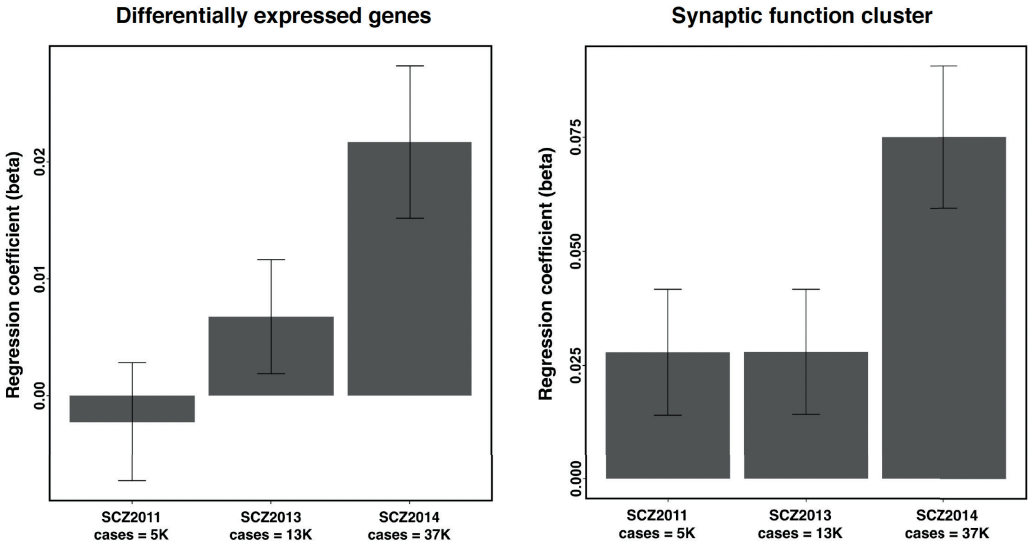


Figure S19. GWAS sample size matters. *(left) A bar plot showing the regression coefficient (MAGMA) of the association between the T2 statistic (likelihood of being differentially expressed) and SCZ gene level test statistics for three SCZ GWAS studies of increasing sample sizes. The numbers of cases for each study are denoted on the x-axis labels. (right) A similar plot showing the association of SCZ risk and membership to the synaptic function cluster for each GWAS. Regression coefficients are shown with corresponding standard errors.*

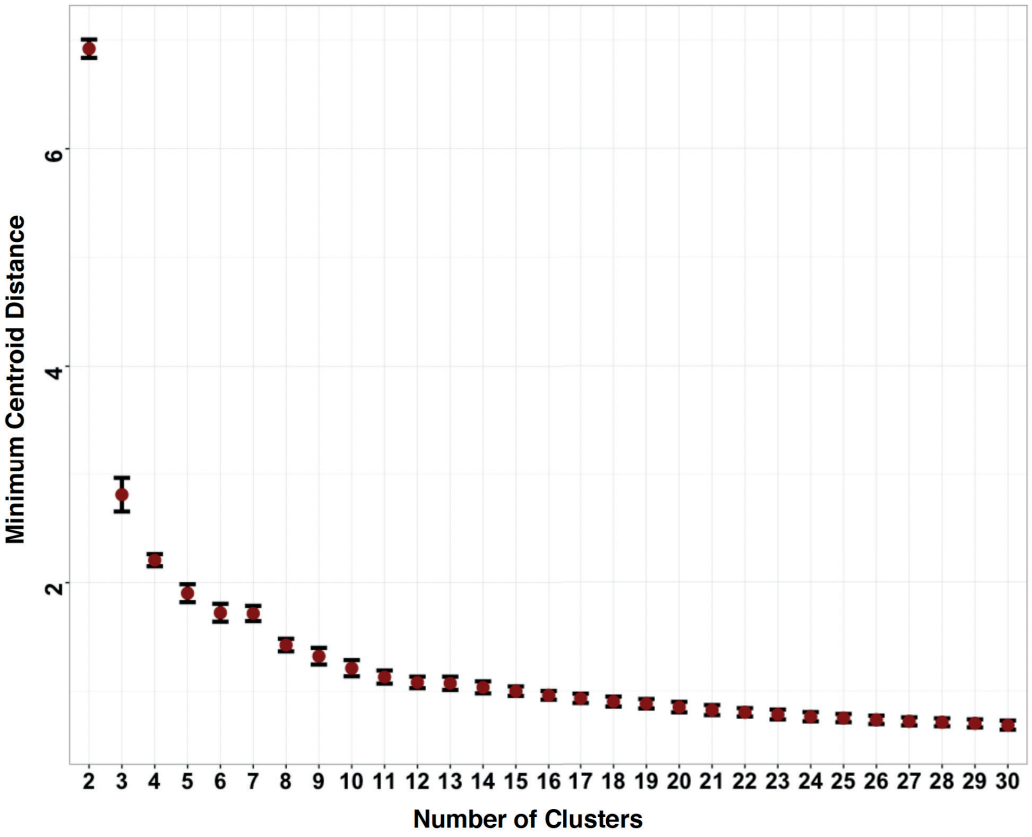


Figure S20. A plot showing minimum centroid distance against increasing numbers of clusters. *We sampled 100 independent single-replicate time series (see supplementary figure 5) and performed fuzzy c-means clustering for each time series across various numbers of clusters with a fuzzifier of 1.55. For each we calculated the minimum centroid distance across clusters. Shown above in red are the mean across time series with corresponding standard errors in black. The x-axis shows the number of clusters and the y-axis the minimum centroid distance. The optimal cluster number is chosen as the number before which there starts a gradual decrease in minimum centroid distance as cluster number increases. This indicates that additional clusters add little information. The optimal cluster number was set at 8.*

Day 0	Day 2	Day 5	Day 10	Day 15	Day 20	Day 30
Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1
Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2
Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3
Replicate 4	Replicate 4	X	X	X	X	Replicate 4

Day 0	Day 2	Day 5	Day 10	Day 15	Day 20	Day 30
Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1	Replicate 1
Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2	Replicate 2
Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3	Replicate 3
Replicate 4	Replicate 4	X	X	X	X	Replicate 4

Figure S21. A schematic example of sampling independent single-replicate time series. *We calculated average cluster membership for each probe for each cluster across 100 independently sampled single-replicate time series. Give the data we can sample 5,184 independent single-replicate time series (4^3 X 3^4). Above are two dummy examples shown of how a single-replicate time series could look like. The yellow color denotes the sampled samples and the red line shows a path that defines the single-replicate time series that these samples make up.*

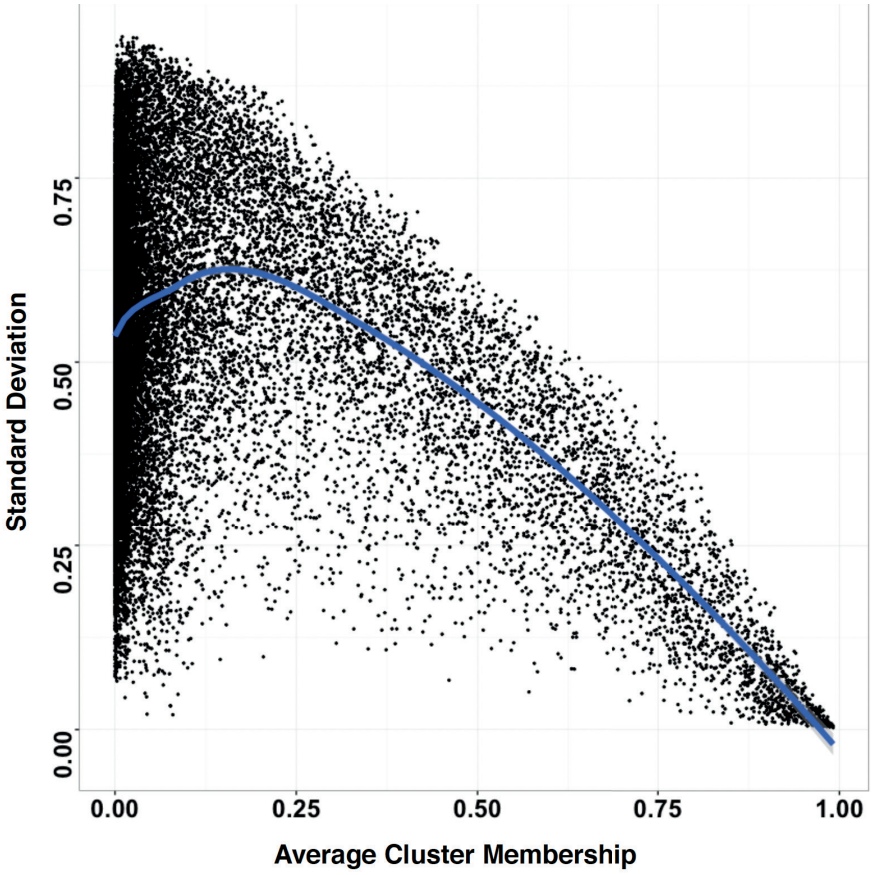


Figure S22. A plot showing the variation in cluster membership values across 100 independently sampled time series. *We performed soft clustering on 7,734 probes using fuzzy c-means clustering with a fuzzifier of 1.55 and a cluster number of 8. Cluster memberships were calculated as the average membership determined across 100 independently sampled time series. The x-axis above shows average cluster membership and the y-axis the standardized standard deviation. Data is shown for a specific cluster with each dot representing a probe. The blue line represents a smoothened curve representing the relationship between standard deviation and average membership with 95% confidence intervals in grey. This relationship is consistent across all 8 clusters.*

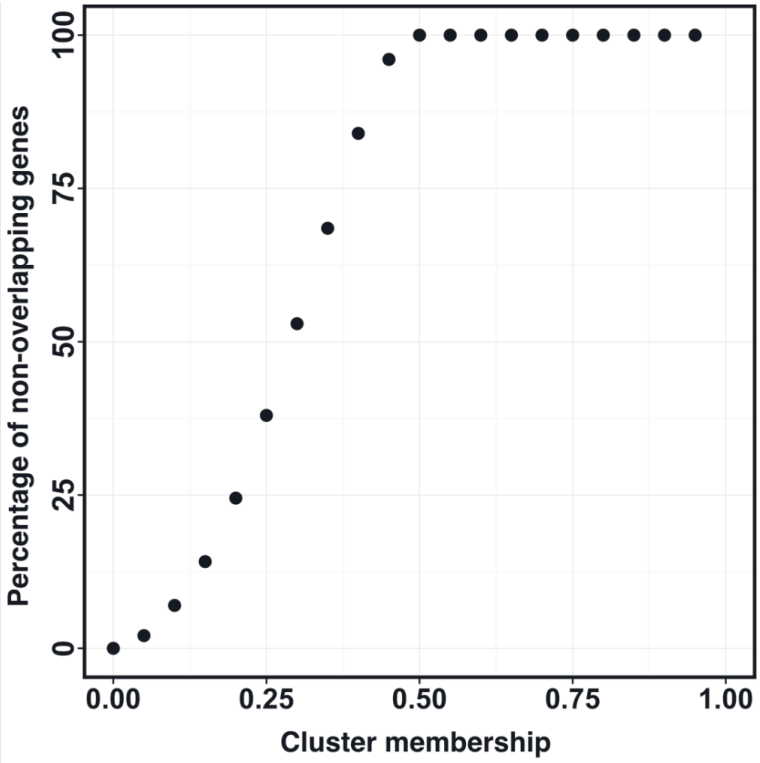


Figure S23. The overlap between clusters across the membership range. *The percentage of unique genes by Ensembl ID was calculated across different membership values for each cluster. These percentages were subsequently averaged across 8 clusters. The y-axis shows the average percentage of unique genes (i.e. no overlap between clusters) with membership value on the x-axis.*

References – supplementary information

1. Fedoroff S, Richardson A (2010): Protocols for Neural Cell Culture, 4th ed. (L. C. Doering, editor). Humana Press, pp 51–73.

2. Shi Y, Kirwan P, Livesey FJ (2012): Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. Nat Protoc. 7: 1836–1846.

3. Zhang X-Q, Zhang S-C (2010): Differentiation of Neural Precursors and Dopaminergic Neurons from Human Embryonic Stem Cells. In: Turksen K, editor. Human Embryonic Stem Cell Protocols. Totowa, NJ: Humana Press, pp 355–366.

4. Amit M, Carpenter MK, Inokuma MS, Chiu CP, Harris CP, Waknitz MA, et al. (2000): Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Dev Biol. 227: 271–278.

5. Thomson JA (1998): Embryonic Stem Cell Lines Derived from Human Blastocysts. Science. 282: 1145–1147.

6. Laurent LC, Nievergelt CM, Lynch C, Fakunle E, Harness JV, Schmidt U, et al. (2010): Restricted ethnic diversity in human embryonic stem cell lines. Nat Methods. 7: 5–6.

7. Ware CB, Nelson AM, Blau CA (2006): A comparison of NIH-approved human ESC lines. Stem Cells. 24: 2677–2684.

8. Langfelder P, Horvath S (2008): WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9: 559.

9. Oldham MC, Langfelder P, Horvath S (2012): Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst Biol. 6:

10. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. (2014): An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. Journal of Neuroscience. 34: 11929–11947.

11. Stein JL, de la Torre-Ubieta L, Tian Y, Parikshak NN, Hernández I a., Marchetto MC, et al. (2014): A quantitative framework to evaluate modeling of cortical development by neural stem cells. Neuron. 83: 69–86.

12. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, et al. (2011): Spatio-temporal transcriptome of the human brain. Nature. 478: 483–489.

13. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, et al. (2014): Transcriptional landscape of the prenatal human brain. Nature. 508: 199–206.

14. Plaisier SB, Taschereau R, Wong JA, Graeber TG (2010): Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. Nucleic Acids Res. 38: e169–e169.

15. Tai YC, Speed TP (2006): A multivariate empirical Bayes statistic for replicated microarray time course data. Ann Stat. 34: 2387–2412.

16. Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J (2009): An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). BMC Bioinformatics. 10: 409.

17. Kumar L, E Futschik M (2007): Mfuzz: a software package for soft clustering of microarray data. Bioinformation. 2: 5–7.

18. Schwämmle V, Jensen ON (2010): A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. Bioinformatics. 26: 2841–2848.

19. Huang DW, Lempicki R a., Sherman BT (2009): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4: 44–57.

20. Lek M, Karczewski KJ, Samocha KE, Banks E, Fennell T, O AH, et al. (2016): Analysis of protein-coding genetic variation in 60,706 humans. Nature. 536: 285–291.

21. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Consortium SWG of TPG, et al. (2015): LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 47: 291–295.

22. CONVERGE Consortium (2015): Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 523: 588.

23. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. (2016): Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nature Publishing Group. 48: 1031–1036.

24. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, et al. (2013): A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 18: 497– 511.

25. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015): A global reference for human genetic variation. Nature. 526: 68–74.

26. de Leeuw CA, Mooij JM, Heskes T, Posthuma D (2015): MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 11. doi: 10.1371/journal.pcbi.1004219.

27. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, et al. (2017): Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat Genet. . doi: 10.1038/ng.3954.

28. Finucane H, Reshef Y, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. (2017): Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. bioRxiv. . doi: https://doi.org/10.1101/103069.

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81: 559–575.

30. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015): Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 4:7.

31. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. (2015): Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 47: 1228–1235.

32. van de Leemput J, Boles NC, Kiehl TR, Corneo B, Lederman P, Menon V, et al. (2014): CORTECON: A temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron. 83: 51–68.

33. Love MI, Huber W, Anders S (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

34. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. (2014): A framework for the interpretation of de novo mutation in human disease. Nat Genet. 46:944–950.

35. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R, et al. (2015): An Atlas of Genetic Correlations across Human Diseases and Traits. Nat Genet. 47:1237–1241.

36. Gunnell D, Harrison G, Whitley E, Lewis G, Tynelius P, Rasmussen F (2005): The association of fetal and childhood growth with risk of schizophrenia. Cohort study of 720,000 Swedish men and women. Schizophr Res. 79: 315–322.

37. Zammit S, Rasmussen F, Farahmand B, Gunnell D, Lewis G, Tynelius P, Brobert GP (2007): Height and body mass index in young adulthood and risk of schizophrenia: A longitudinal study of 1 347 520 Swedish men. Acta Psychiatr Scand. 116: 378–385.

38. Bacanu SA, Chen X, Kendler KS (2013): The genetic overlap between schizophrenia and height. Schizophr Res. 151: 226–228.

39. Tanapat P (2013): Neuronal Cell Markers. Materials and Methods. 3. doi:10.13070/mm.en.3.196.

40. Magavi SSP, Macklis JD (n.d.): Immunocytochemical Analysis of Neuronal Differentiation. Neural Stem Cells. pp 291–298.

41. von Bohlen Und Halbach O (2007): Immunohistological markers for staging neurogenesis in adult hippocampus. Cell Tissue Res. 329: 409–420.

CHAPTER 4

Integrative genomic strategies applied to a lymphoblast cell line model reveal specific transcriptomic signatures associated with clozapine response

Authors

S.A.J. (Jytte) de With[#,1,2]
Anil PS Ori[#,1]
Tina Wang[1]
Sara L Pulit[3]
Eric Strengman[1]
Joana Viana[4]
Jonathan Mill 5
Simone de Jong[6]
Roel A Ophoff[1,7,8]

Affiliations

[1.] UCLA Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, California, USA
[2.] Brain Center Rudolf Magnus, Department of Psychiatry, University Medical Center Utrecht, The Netherlands
[3.] Department of Medical Genetics, University Medical Center Utrecht, The Netherlands
[4.] Institue of Psychiatry, SGDP Research Centre, King's College London, United Kingdom
[5.] Uiversity of Exeter Medical School, University of Exeter, Exeter, United Kingdom.
[6.] MRC Social, Developmental and Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
[7.] Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA.
[8.] Department of Psychiatry, Erasmus Medical Center, Erasmus University, Rotterdam, The Netherlands.

[#]These authors contributed equally

Abstract

Clozapine is an important antipsychotic drug. However, its use is often accompanied by metabolic adverse effects and, in rare instances, agranulocytosis. The molecular mechanisms underlying these adverse events are unclear. To gain more insights into the response to clozapine at the molecular level, we exposed lymphoblastoid cell lines (LCLs) to increasing concentrations of clozapine and measured genome-wide gene expression and DNA methylation profiles. We observed robust and significant changes in gene expression levels due to clozapine (n = 463 genes at FDR < 0.05) affecting cholesterol and cell cycle pathways. At the level of DNA methylation, we find significant changes upstream of the LDL receptor, in addition to global enrichments of regulatory, immune and developmental pathways. By integrating these data with human tissue gene expression levels obtained from the Genotype-Tissue Expression project (GTEx), we identified specific tissues, including liver and several tissues involved in immune, endocrine and metabolic functions, that clozapine treatment may disproportionately affect. Notably, differentially expressed genes were not enriched for genome-wide disease risk of schizophrenia or for known psychotropic drug targets. However, we did observe a nominally significant association of genetic signals related to total cholesterol and low-density lipoprotein levels. Together, these results shed light on the biological mechanisms through which clozapine functions. The observed associations with cholesterol pathways, its genetic architecture and specific tissue effects may be indicative of the metabolic adverse effects observed in clozapine users. LCLs may thus serve as a useful tool to study these molecular mechanisms further.

Manuscript status: submitted
Preprint available: https://doi.org/10.1101/2020.09.22.308262

Introduction

Antipsychotic drugs (APs) play an important role in the treatment of psychotic disorders such as schizophrenia (SCZ). Clozapine is one of the most effective antipsychotic drugs (AP) (Leucht et al. 2013; Kane et al. 1988; Taylor 2017). However, the decision to prescribe clozapine is complicated by its potential to induce severe adverse effects (Leucht et al. 2013). The most severe adverse effect, with a prevalence of <1%, is clozapine-induced agranulocytosis, a dramatic reduction of white blood cells (Andersohn, Konzen, and Garbe 2007). More common adverse effects include weight gain, dyslipidemia and type 2 diabetes. These adverse metabolic effects are, in addition to the chance of developing agranulocytosis, the primary reasons for patient noncompliance and discontinuation of treatment (Cohen 2014; Weiden, Mackell, and McDonnell 2004).

The biological mechanisms underpinning the effect of clozapine, as well as its adverse effects, remain elusive. A twin study estimated that the heritability of APs-induced weight gain is approximately 60%(Gebhardt et al. 2010), suggesting a substantial role for genetic factors. Candidate genes studies of clozapine-induced adverse effects have yielded ambiguous results and lack consistent replication (reviewed by(Roerig, Steffen, and Mitchell 2011; Müller, Chowdhury, and Zai 2013; Chowdhury, Remington, and Kennedy 2011; Lett et al. 2012; Yan, Chen, and Zheng 2013)). Two genome-wide association studies (GWAS) investigating antipsychotic-induced metabolic adverse effects have yielded inconclusive findings, primarily due to insufficient sample sizes and the potential polygenic nature of this trait (Malhotra et al. 2012; Adkins et al. 2011).

Intermediate molecular phenotypes, such as gene expression studies in specific cell lines or tissues, may improve our understanding of the molecular function of clozapine. Previous studies have found that atypical antipsychotic drugs may induce cholesterol metabolism through transcription factors such as sterol regulatory element binding proteins (SREBP1 and 2) (Ferno et al. 2011), suggesting that drug-induced cholesterol metabolism is related to these adverse metabolic effects. However, such findings were not consistently replicated when profiling whole blood (Ferno et al. 2011; Harrison et al. 2016; Vik-Mo et al. 2008). It is possible that changes in DNA methylation could mediate changes in gene expression, such as the AP-induced hypomethylation of the FAR2 gene leading to insulin resistance (Burghardt et al. 2016). However, such studies are currently limited and those performed provide inconsistent results (Swathy et al. 2017; Swathy and Banerjee 2017; Stapel et al. 2017; Ota et al. 2014; Melas et al. 2012; Kinoshita et al. 2017; Burghardt et al. 2016; Houtepen et al. 2016; Rukova et al. 2014).

A major obstacle towards understanding clozapine-induced metabolic effects is that clozapine therapy is a relatively rare (~6%) treatment plan in patients diagnosed with schizophrenia (Burghardt et al. 2016; Stroup et al. 2016), of which then only 1% develop CIA. Such factors challenge our ability to adequately sample a large and controlled prospective cohort of patients thereby limiting progress in understanding both metabolic and hematological adverse effects. To augment the lack of available *in vivo* data, we implemented an *in vitro* lymphoblast cell line (LCL) model to study the effects of drug exposure at the molecular level. Cell-based models have been successfully employed for pharmacogenomic studies, including LCL models to study clozapine function (Welsh et al. 2009; Wen et al. 2012; Morag et al. 2010; de With et al. 2015). Here, we exposed LCLs to increasing doses of clozapine and collected both expression

and methylation profiles. Through integrative genomic analyses, we aim to extrapolate *in vitro* molecular signatures of clozapine towards relevance of *in vivo* function and study clozapine response without the need to assemble a large cohort of patients. We identified significant changes in transcriptomic signatures associated with cholesterol and cell cycle pathways. By then integrating these molecular profiles with genome-wide association study (GWAS) summary statistics of different traits and diseases, we show that clozapine-associated genes also overlap with genetic signals related to total cholesterol and low-density lipoprotein (LDL) levels but not schizophrenia genetic risk. Clozapine-associated genes are furthermore related to specific human tissues, such as liver and those involved in immune, endocrine and metabolic functioning.

Methods

Lymphoblast cell lines

We used lymphoblast cell lines (LCLs) from four unrelated samples, all part of the collection of Utah residents of Northern and Western European ancestry (HapMap CEPH/CEU phase 1) (Consortium and †The International HapMap Consortium 2003). We obtained LCLs from the Coriell Institute for Medical Research (Camden, NJ, USA) and maintained the cell lines as previously described (Consortium and †The International HapMap Consortium 2003; de With et al. 2015). To study methylation changes after exposure to clozapine, we performed a separate experiment using six LCLs (HapMap, CEPH/CEU phase 1) consisting of two parent-offspring trios.

In vitro experimental design and clozapine exposure

Clozapine, purchased from Sigma Aldrich, was dissolved in culture medium with dimethyl sulfoxide (DMSO), with a maximum concentration of 0.1%. Clinical concentration of clozapine was set at 2μM (Baumann et al. 2004). To enhance the downstream molecular effects of clozapine-exposure, we chose to expose the lymphoblast cells with supratherapeutic concentrations, as was done in previous studies of clozapine (Leykin, Mayer, and Shinitzky 1997; Tschen et al. 1999; de With et al. 2015). Cell lines were exposed for 24 hours to clinical concentration (Supplementary Methods), 10x, 50x and 100x clinical concentration (20μM-100 μM-200 μM clozapine) and vehicle (DMSO); each concentration was measured in 4 cell lines, after which RNA was obtained for gene expression analysis (Supplemental Figure 1A). To study DNA methylation changes in response to clozapine, we subsequently performed an independent experiment similar to the gene expression experiment. LCLs were exposed to vehicle DMSO, 1x, 20x, 40x and 60 times clinical concentration for 24h and 96h (Supplemental Figure 1B). We measured cell viability using the TC10 automated cell counter.

Sample processing and data collection

After desired exposure time, we lysed cells and performed RNA and DNA collections using column-based extraction methods from Qiagen according to manufacturer's instructions (Supplementary Methods). Gene expression profiling was carried out using Illumina® HumanHT-12 v4 Expression BeadChip technology. DNA methylation assays were performed with Illumina® Infinium HumanMethylation450 Beadchip arrays.

Data preprocessing and normalization

We processed raw gene expression values using the "Lumi" R-package (Du, Kibbe, and Lin 2008). We log2 transformed and quantile normalized the raw data, keeping only expressed gene transcripts (detection $p < 0.01$) for further analysis (22,926 probes). We processed DNA methylation values using the "WateRmelon" Bioconductor package (Pidsley et al. 2013), removing probes that were known to cross-hybridize, probes containing SNPs in target CpG regions, probes with detection p-value greater than 0.01 in 5% of samples, and probes with beadcounts > 3 (n=86,068 probes in total)(Chen et al. 2013; Price et al. 2013). We normalized the data using the dasen function and computed β-values, defined as the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities), to measure methylation levels. To limit the effect of heteroskedasticity, we included only variable probes with β-values between 0.2 – 0.8 in our analyses (165,014 probes) (Du et al. 2010).

Statistical analyses

To detect clozapine-induced molecular changes, for each probe, we tested for association between gene expression (1) or DNA methylation levels (2) with increasing clozapine concentrations using the following linear models implemented in R using the "Limma" package (Smyth, n.d.):

$$Y_{g,i} = \beta 0_{g,i} + \beta 1_{g,i} Cloz + \beta 2_{g,i} RIN + \varepsilon_{g,i} \quad (1)$$

$$Y_{m,i} = \beta 0_{m,i} + \beta 1_{m,i} Cloz + \varepsilon_{m,i} \quad (2)$$

where Y is the normalized gene expression (g) or DNA methylation levels (m) for an individual probe, b0 the intercept, b1 the effect of clozapine concentration, b2 the effect of RIN, and e the residual variation of the model. We ran each model for each probe per individual i and subsequently performed a meta-analysis across all individuals by combining p-values using Stouffer's method with directionality of the effect sizes taken into account. We included RNA-integrity number (RIN) as a covariate in the gene expression model and applied a Bonferroni correction to correct for multiple testing, resulting in a significance threshold of $p < 2.18 \times 10^{-6}$ for the gene expression analysis (n = 22,926 probes) and $p < 3.03 \times 10^{-7}$ for the DNA methylation analysis (n = 165,014 probes).

Gene ontology analysis

We performed functional gene ontology analysis using DAVID (Database for Annotation, Visualization and Integrated Discovery, version 6.8, interrogated February 2018)(Huang, Lempicki, and Sherman 2009; Huang, Sherman, and Lempicki 2009), with default settings (Supplementary Methods).

Functional enrichment analysis of DNA methylation data

We used Genomic Regions Enrichment of Annotations Tools (GREAT, v3.0) to predict the biological function of the top methylation probes associated with clozapine exposure. GREAT links both proximal and distal genomic CpG sites with their putative target genes and implements both a gene-based test and a region-based test using the hypergeometric and binomial test, respectively, which allows us to assess enrichment of genomic regions in biological annotations of pathway databases (Supplementary Methods)(McLean et al. 2010). The statistical outputs of GREAT for both gene-based and region-based tests were subsequently adjusted for multiple testing using Bonferroni correction.

Weighted gene co-expression network analysis

We performed a gene expression network analysis using weighted gene co-expression network analysis (WGCNA) in R. Briefly, WCGNA identifies distinct modules using the shared variation in gene expression based on pairwise correlation. To account for the biases related to differing probe numbers between genes assayed on the array, we provided as input the mean probe expression of genes residing within nominally significant differentially expressed genes ($p<0.05$), considering 5,708 probes within 4,897 genes (Horvath 2011; Langfelder and Horvath 2008; Zhang and Horvath 2005). To assign biological function to each WCGNA module, we performed gene ontology analysis using DAVID we performed.

Additional DNA methylation analyses

We ran a candidate gene study for CpG sites in close proximity of genes with evidence of clozapine-induced differential gene expression. Methylation probes within the gene body, in the 3' and 5' untranslated regions and up to 1,500 nucleotides upstream of the transcription start site of the 463 'top genes' ($p<0.05$) were selected for a post-hoc analysis ($n = 1,004$). These results can be found in Table 1 of the Supplemental Material.

GTEx cross tissue analysis

To translate the *in vitro* effects of clozapine to *in vivo* human biology, we investigated how clozapine genes behave across 22 human tissues represented in the GTEx data set. We downloaded gene level quantifications (version 6, date: April 24, 2019) from the GTEx Project web portal (Melé et al. 2015) and transformed gene expression values using a log2 transformation (1 + RPKM value). We then tested 1) if clozapine-associated genes have higher or lower average expression within each tissue and 2) if between tissue distance is different for clozapine-associated genes compared to the expected based on chance. To calculate between tissue distance, we used the top half most variable genes across GTEx samples that were also significantly detected in our *in vitro* assay (n=7,025). We then performed multidimensional scaling using the isoMDS() function in the MASS R package (v7.3)(Melé et al. 2015; Venables and Ripley 2002) and calculated between tissue Euclidean distances using the dist() function in R (v.3.3.3) and the median gene expression values for each tissue for clozapine associated genes (N=463).

To established whether clozapine-associated genes significantly deviated from the expected, we established a null distribution by repeating this procedure using the Euclidean

distance for samplings of similarly sized sets of genes. Within each sampling, genes were sampled with a probability that matched the distribution of average gene expression of clozapine-associated genes in our LCL experiment. This was done to account for differentially expressed genes having higher gene expression levels. Our null distribution therefore better captured the expected effect. Null distributions were built separately for the expected average gene expression within each tissue and for the expected Euclid distance for each tissue pair. P-values were then established by calculating the proportion of samplings that were higher or lower than the statistic of the observed clozapine-associated genes and corrected for multiple testing by Bonferroni correction. Analyses were conducted separately for genes upregulated and downregulated after clozapine exposure.

MAGMA gene-set analysis

To investigate if clozapine-associated genes were enriched for schizophrenia or cardiovascular-related genetic association signals, we performed gene-set analysis using MAGMA (multimarker analysis of genomic annotation) (de Leeuw et al. 2015). We tested the following three gene-sets, each with three different significance thresholds:

1. All genes differentially expressed, either up- or downregulated;
2. Upregulated differentially expressed genes;
3. Downregulated differentially expressed genes.

We obtained GWAS summary statistics for schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), body mass index (Yengo et al., n.d.), coronary artery disease (Nikpay et al. 2015), type 2 diabetes(Scott et al. 2017), total cholesterol, triglycerides, high-density-lipoprotein, and low-density-lipoprotein(Scott et al. 2017; Surakka et al. 2015). For each GWAS, we computed aggregate gene-level test statistics using a 10kb window around the transcription start and end site of each gene; a total of 11,533 genes were available for analysis. We then tested for association between GWAS gene level Z-scores and experimental gene-sets (defined above). We estimated linkage disequilibrium (LD) using the 1000 Genomes European reference panel (Scott et al. 2017; Surakka et al. 2015; 1000 Genomes Project Consortium et al. 2015) and ran MAGMA using a two-sided competitive test by accounting for probes tested in our experiment while also correcting for gene size, SNP density, minor allele count and LD.

Antipsychotic drug gene-sets and enrichment of schizophrenia heritability

As a complementary analysis, we investigated whether manually curated antipsychotic drug gene-sets overall are associated with schizophrenia heritability. We performed gene-set analysis on antipsychotic drug targets that were previously reported to be enriched for schizophrenia heritability [55]. Drug target list were curated using drug-gene interaction databases (Wagner et al. 2016; Roth et al. 2000). We identified 50 sets of drugs with ATC (Anatomical Therapeutic Chemical) code N05A, a drug class to which all antipsychotic drugs belong, including clozapine. Using the schizophrenia GWAS as input to MAGMA, we ran gene-set analysis (1) per individual N05A antipsychotic drug gene-set, (2) for all N05A drug target genes combined, and for (3) all N05A drug target genes combined excluding clozapine.

LD Score Regression estimation of heritability and genetic correlations

Pre-computed LD scores from the 1000 Genomes European reference panel provided through the LD Score Regression (LDSR) github and GWAS summary statistics file processed and reformatted to sumstats format were used as input to (LDSR) to estimate SNP-based heritability (SNP-h^2) (B. K. Bulik-Sullivan et al. 2015). Genetic correlations between traits were estimated using cross-trait LDSR via the --rg flag (B. Bulik-Sullivan et al. 2015).
110

Results

Clozapine exposure induces widespread gene expression changes

We exposed lymphoblast cell lines to increasing concentrations of clozapine and used a linear regression model to identify genes with a subsequent dose-response change in expression levels (Supplemental Figure 1A). In total, we tested 22,926 gene expression probes, of which 5,708 showed nominal significance ($p<0.05$) and 518 probes exceeded a Bonferroni-corrected $p < 2.18*10^{-6}$. These 518 probes consisted of 234 up-regulated probes and 284 down-regulated probes, representing a total of 463 unique genes, which we define as our main set of genes robustly associated with clozapine exposure. The top 10 up-regulated and down-regulated probes are shown in Table 1. See Supplement for the complete list of differentially expressed genes.

Up-regulated genes		Down-regulated genes	
Gene	P-value	Gene	P-value
STMN1 (Stathmin 1)	$5.16^{*}10^{-16}$	*LSS* (Lanosterol synthase)	$6.14^{*}10^{-17}$
HIST2H2AC (Histone cluster 2 H2A family member C)	$1.98^{*}10^{-15}$	*MAL* (Mal, T-cell differentiation protein)	$1.10^{*}10^{-16}$
RPS7 (Ribosomal protein S7)	$2.52^{*}10^{-15}$	*MIR1974* (MicroRNA 1974)	$5.56^{*}10^{-16}$
HIST1H2BJ (Histone cluster 1 H2B family member J)	$4.42^{*}10^{-15}$	*LDLR* (Low density lipoprotein receptor)	$3.97^{*}10^{-14}$
HIST2H2AA3 (Histone cluster 1 H2A family member A3)	$1.97^{*}10^{-14}$	*DHCR7* (7-dehydrocholesterol reductase)	$4.26^{*}10^{-14}$
RPS15 (Ribosomal protein S15)	$2.17^{*}10^{-14}$	*PASK* (PAS Domain Containing Serine/Threonine Kinase)	$5.09^{*}10^{-14}$
HIST2H2AA4 (Histone cluster 2 H2A family member A4)	$2.78^{*}10^{-14}$	*RGS1* (Regular Of G Protein Signaling 1)	$2.31^{*}10^{-13}$
HIST1H2AC (Histone cluster 1 H2A family member C)	$2.79^{*}10^{-14}$	*LYPD6B* (LY6/PLAUR Domain Containing 6B)	$4.38^{*}10^{-13}$
AURKA (Aurora Kinase A)	$5.19^{*}10^{-14}$	*TPP1* (Tripeptidyl Peptidase 1)	$6.77^{*}10^{-13}$
SGOL1 (Shugoshin 1)	$2.09^{*}10^{-13}$	*RENBP* (Renin Binding Protein)	$7.54^{*}10^{-13}$

Table 1. Top 10 up- and downregulated genes after clozapine exposure (previous page). *Gene symbols are shown with corresponding gene name and p-value of differential gene expression analysis. Full list of genes are shown in Supplementary Table 1.*

Gene ontology enrichment analysis of up-regulated genes showed significant functional enrichment for cholesterol metabolism (13 genes, $p = 4.15*10^{-15}$) and steroid biosynthesis (11 genes, $p = 1.01*10^{-8}$) (Table 2), while analysis of down-regulated genes were enriched for cell division processes and related annotations, such as mitosis (44 genes, $p = 1.87*10^{-39}$), chromosome (49 genes $p = 3.03*10^{-35}$) and nucleosome (16 genes, $p = 3.12*10^{-14}$) and other cell cycle pathways (Table 2).

Gene expression network analysis after clozapine exposure

WGCNA network analysis of clozapine-induced differentially expressed genes (N=4,987) yielded 15 co-expression modules, ranging in size from n=61 to 1,791 genes. Five gene co-expression modules were altered upon clozapine exposure: M14 (upregulated, 61 genes), M10 (upregulated, 155 genes), M9 (upregulated, 158 genes), M3 (upregulated, 579) and M1 (downregulated, 1,791 genes), which were nominally significant. The M14 co-expression module was enriched for genes involved in cholesterol metabolism (13 genes, $p = 4.8*10^{-15}$), the M1 co-expression module was enriched for genes involved in cell cycle (133 genes, $p = 7.3*10^{-32}$) and the M9 and M3 co-expression module was enriched for mitochondrial genes (15 genes, $p = 6.7*10^{-6}$ and 46 genes $p = 3.8*10^{-7}$ respectively). The M10 co-expression module was enriched for genes involved in the nucleosome (9 genes, $p = 1.4*10^{-7}$).

Minimal changes in DNA methylation at single CpG sites after clozapine exposure

We then performed DNA methylation profiling to assess whether these effects were due to epigenetic changes. For statistical analysis, we applied an analytical approach similar to our gene expression analyses. Targeted analysis on the 1,004 CpG sites near the 463 differentially-expressed genes revealed 3 probes exhibiting significant changes in DNA methylation ($p < 1.08*10^{-4}$), including a probe upstream of the low-density lipoprotein receptor (LDL-R) gene (cg22971501, $p=4.75*10^{-5}$) after 24h; one probe upstream of the cyclin F (CCNF) gene showed a significant change in DNA methylation after 96h (Table S4). Beyond these examples, global methylation differences were not observed after 24h or 96h at the level of individual probes.

Functional annotation cluster	Enrichment score	Best P-value	Best fold enrichment	Average P-value	Average fold enrichment	Number of pathways
Upregulated pathways						
Cholesterol/lipid/steroid biosynthesis	8.44	$4.15^{*}10^{-15}$	46.92	$1.19^{*}10^{-3}$	18.46	19
Cholesterol/Steroid biosynthesis	4.77	$1.01^{*}10^{-8}$	68.45	$4.66^{*}10^{-4}$	54.32	3
Down-regulated pathways						
Mitosis/cell division	31.26	$1.87^{*}10^{-37}$	14.77	$1.12^{*}10^{-21}$	10.69	5
Chromosome/centromere	16.47	$3.03^{*}10^{-35}$	21.32	$7.76^{*}10^{-8}$	15.79	9
Histone/nucleosome	6.93	$3.12^{*}10^{-14}$	20.72	$1.73^{*}10^{-3}$	9.13	18
Spindle	6.02	$4.95^{*}10^{-12}$	12.89	$7.45^{*}10^{-4}$	10.13	6
Nucleotide/ATP-binding	5.72	$1.46^{*}10^{-13}$	2.71	$7.78^{*}10^{-5}$	2.47	5
Microtubule/kinesin	4.60	$1.90^{*}10^{-4}$	16.92	$4.99^{*}10^{-2}$	8.56	19

Table 2. Main functional annotations associated with clozapine exposure based on gene expression. *Output of pathway enrichment analyses using DAVID is shown. The enrichment score is used as the main metric of importance. It is defined as the geometric mean of all enrichment p-values of each annotation term within the group. It is expressed as the minus log of the p-value, an enrichment score of 1.3 is nominally significant. This table shows clusters with an enrichment score > 4.6, corresponding to a p-value < 0.01. Fold enrichment is a measure to express the enrichment of this particular group of genes in comparison with the genes in the human genome. A list of all pathways is available in supplemental information.*

DNA methylation is affected at the pathway level

To investigate if our top associated DNAm probes aggregated to changes at the pathway level, we performed pathway enrichment analysis using GREAT, which incorporates functional annotation from various databases to predict cis-regulatory function of genomic regions of interest. When considering probes exhibiting FDR < 10%, we observed enrichment for protein binding and regulation of cellular processes after 24h of clozapine treatment. When considering the top 1000 probes, we also observed enrichment of immune-related functions, such as the Major Histocompatibility Complex (MHC) class II protein complex and Graft- versus-host disease after 24h. We found significant enrichment for estradiol regulation and various embryonic developmental processes (Figure 1 and Table S5) when considering the top 1000 probes after 96h.

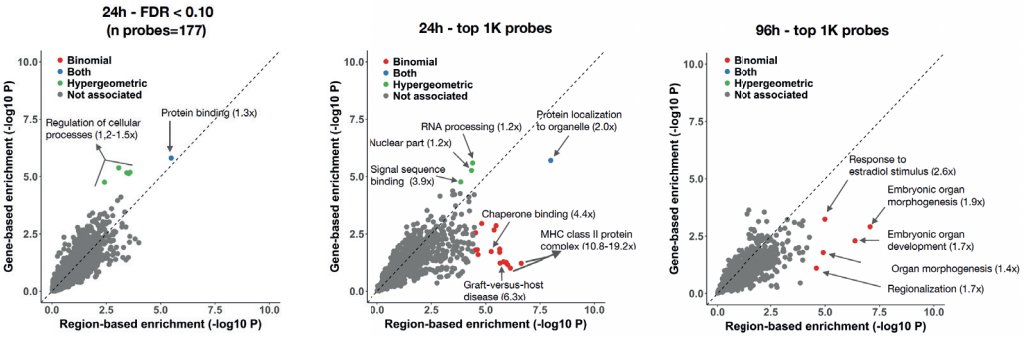


Figure 1. Clozapine-associated DNAm probes concentrate to specific biological annotations (previous page). *Genomic Regions Enrichment of Annotations Tool (GREAT) was used to assign biological context to genomic regions associated to clozapine exposure. GREAT uses the annotation of nearby genes and regulatory elements to performed enrichment analysis across known functional databases. The x-axis shows the -log 10 p-value of the region-based analysis (binomial test), while the y-axis shows the enrichment for the gene-based analysis (hypergeometric overlap test). Results are shown for (A) the top probes after 24 hours (FDR < 0.10, n=177), (B) the top 1,000 probes after 24 hours, and (C) the top 1,000 probes after 96 hours. Each dot represents an annotation. Significant annotations, after Bonferroni correction, are color-coded according to the test used.*

DNA methylation changes are time-dependent

To examine if clozapine affected each time point similarly, we correlated effect sizes at each CpG probe between time points. Across all probes, we found a significant negative correlation between the Z-score of the association between DNA methylation levels and clozapine exposure at 24h and 96h (Pearson r = -0.32, $P<2.2x10^{-16}$). This correlation was preserved among probes at FDR < 10% at 24h (n=177, Pearson r = -0.38, $P=1.3x10^{-7}$) but not among the top probes at 96 hours (n=177, Pearson r = -0.12, P = 0.10), highlighting possible time-dependent effects after clozapine exposure (Figure 2).

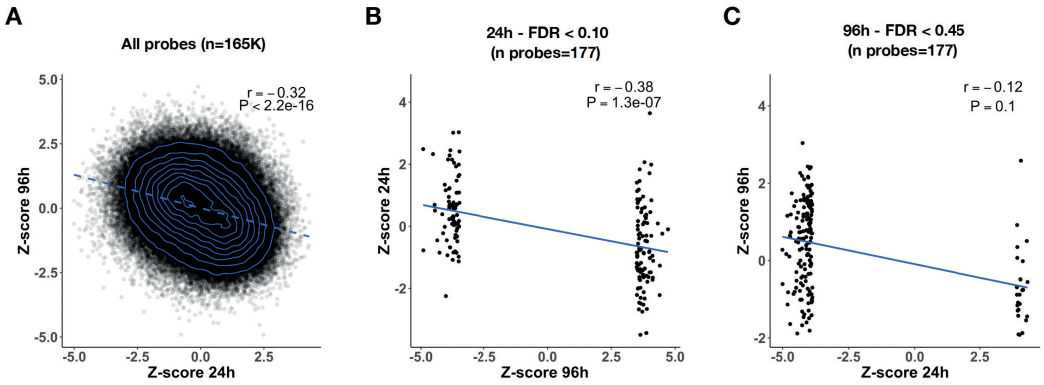


Figure 2. Clozapine-induced DNA methylation changes are time-dependent. *Correlation analysis was performed between DNAm probe associations after 24 and 96 hours of clozapine exposure. The Pearson correlation was computed for (A) all probes tested, (B) the top probes associated after 24 hour exposure (FDR < 0.10, n = 177 probes), and (C) the top 177 probes after 96 hours of drug exposure.*

Clozapine transcriptomic profiles highlight multi-tissue effects in GTEx

We then asked whether our *in vitro* derived transcriptomic signatures could be used to help translate the function of clozapine in humans. For this purpose, we used gene expression data from the GTEx Project, including LCLs (n = 22 GTEx tissues, Figure 3A). First, we overlapped preferentially-expressed genes of each GTEx tissue, as previously reported (Melé et al. 2015), with the clozapine-associated genes detected in our assay. Preferentially expressed genes in GTEx-LCLs exhibited the most overlap with the genes identified by our experiment followed by whole

blood (Supplementary Figure 2). We then investigated if the mean tissue expression of clozapine-associated genes was significantly different from the mean expression of all the genes tested in our experiment. We found that genes downregulated by clozapine have lower expression in all tissues except testis and LCLs (Figure 3B and Table S9). Downregulated genes are enriched for cell cycle processes and their higher expression in the testis and LCLs likely represent the proliferative nature of these tissues. Genes upregulated by clozapine have significantly higher average expression in liver, muscle, lung, and fibroblasts, and lower average expression in testis tissue. We then asked whether clozapine-associated genes have different between-tissue distances as a proxy to investigate possible functional links with other tissues. Distance is calculated using the Euclidean distance measure. A lower value indicates that two tissues are more similar while a higher distance value indicates that these tissues are more dissimilar. Of the 406 tissue pairs, we found that genes upregulated by clozapine have significantly deviating between-tissue distances across 31 pairs of tissues (Figure 3C). Spleen tissue stands out with having the most different distance with other tissues (N=20). We also found multiple differences for adipose, lung, and breast tissue pairs. The distance between cervix uteri and ovary tissue, adrenal gland and thyroid tissue, muscle and nerve tissue are significantly more dissimilar as well for upregulated clozapine genes compared to all genes detected in our assay. For downregulated clozapine genes, we detected significantly dissimilar distances for 27 tissue-pairs, all involving testis tissue (Figure 3C).

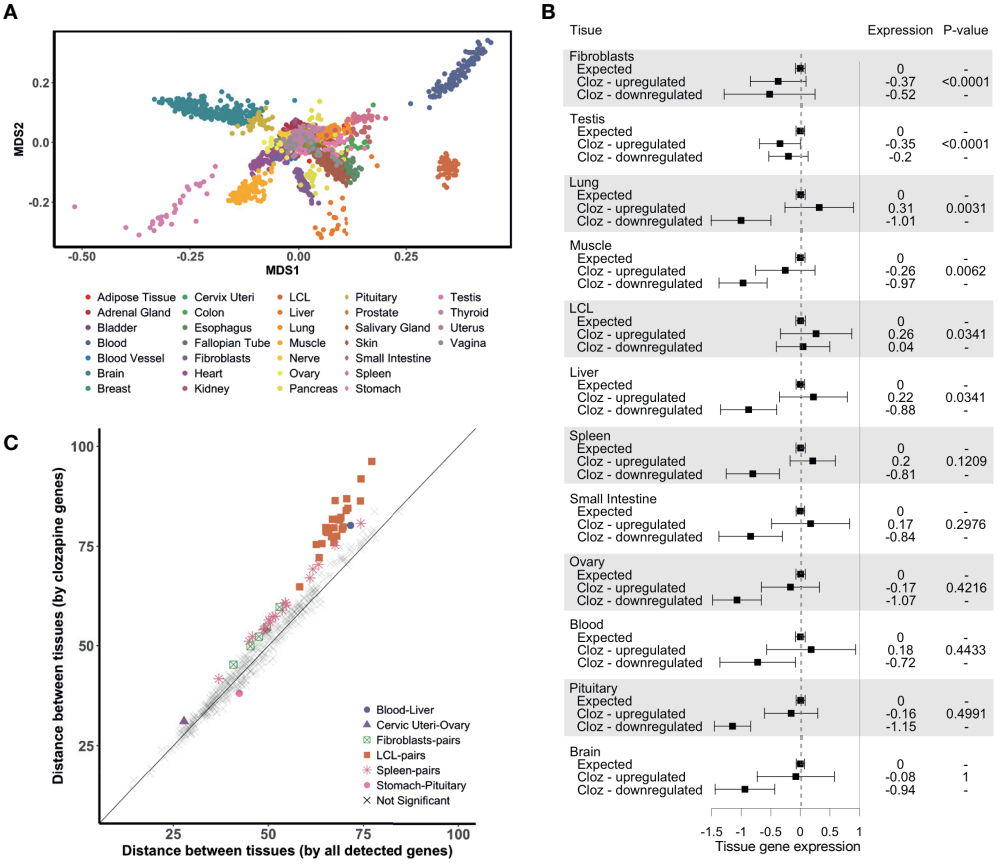


Figure 3. Clozapine-associated genes show tissue-specific expression patterns in GTEx (previous page). *(A) Using gene expression data from the GTEx project, multidimensional scaling (MDS) was performed to visualize tissues and their relationships in a parsimonious way. The top half most variable genes (n=7,025), that were also significantly detected in our experiment, across a random subset of 2,000 GTEx samples were used as input in the MDS analysis. (B) A forest plot visualizing mean tissue gene expression of clozapine-associated genes across several tissues. Clozapine genes were divided into groups that are up regulated and down regulated after drug exposure. The expected mean gene expression, based on 10,000 weighted samplings, is shown as well (dotted vertical line). Within each tissue, the observed mean tissue gene expression (x-axis) is normalized by subtracting the expected mean gene expression. P-values indicate whether the mean gene expression of clozapine up regulated genes significantly deviate from the expected mean expression within a tissue. P-values are corrected for the number of tissues tested. Downregulated genes were not tested. See Supplementary Tables for results of all tissues (C) Between tissue distance across all GTEx tissue pairs. Each point represents one tissue pair. The y-axis shows the Euclidean distance between tissues computed using only clozapine-associated genes. The x-axis shows the expected mean Euclidean distance across 50,000 weighted samplings. Tissue pairs for which the between tissue distance, based on clozapine genes only, significantly deviates from mean expected distance (Padjusted <0.05), are color-coded. P-values are adjusted for multiple testing by Bonferroni correction (n test = 465 pairs x tests = 930).*

Clozapine transcriptome signatures are not enriched for schizophrenia disease risk

Previous studies have found associations between schizophrenia genetic susceptibility and antipsychotic drug targets (Gaspar and Breen 2017; Skene et al. 2018). To investigate whether clozapine-induced *in vitro* gene expression signatures are also associated with genetic susceptibility of schizophrenia, we conducted gene-set enrichment analysis using all differentially expressed genes (n=463). We did not observe a significant enrichment ($p = 0.91$). We then considered upregulated genes, and did not observe significant enrichment ($p = 0.74$), nor for downregulated genes ($p = 0.64$). Gene sets defined according to <1% and <5% FDR did not change these findings (Table S7).

To further understand these findings, we investigated whether SCZ genetic susceptibility aggregates to antipsychotic drug target genes with and without clozapine targets. As SCZ genetic risk has been associated with antipsychotic drug target genes, it could be that this signal is primarily driven by non-clozapine genes. To examine this, we used drug target gene lists from drug-gene interaction databases as previously reported (Wagner et al. 2016; Roth et al. 2000). We were able to extract 53 drug target gene-sets belonging to the N05A class of drugs with antipsychotic actions, including clozapine. Across drug target gene-sets, we mapped all targets to 104 unique genes of which 41 were detectable in our gene expression data but none overlapped with differentially expressed genes identified. We found no evidence for SCZ risk to be enriched in antipsychotic drug target genes overall (n = 96 genes, $p = 0.96$) nor with clozapine targets excluded (n = 52, genes, $p = 0.60$). We did observe a strong concordance between the p-values of individual drug gene-sets tested in our analysis and the p-values reported by the previous study (Gaspar and Breen 2017) (n = 50 drug gene sets, rho = 0.85, $p = 1.49*10^{-15}$), indicating our analysis framework was able to reproduce previous findings at the level of individual drug target gene-sets. Our analysis however does not observe an association between schizophrenia genetic risk and clozapine-associated genes nor antipsychotic drug target genes.

Genes upregulated after clozapine exposure show an association with GWAS signal of total cholesterol and low-density lipoprotein.

Next, we explored whether differentially expressed genes were enriched for genetic signal of cholesterol- and cardiovascular disease-related traits. Using summary statistics of large GWASs for each trait, we observe significant SNP-h^2 and a rich correlation structure between these traits (Figure 4A and 4B). We then integrated the observed SNP-h^2 with our detected clozapine transcriptomic signatures. While no association remained significant after correction of multiple testing (72 tests, $p < 6.9*10^{-4}$), we did observe an increasing association between total cholesterol and LDL heritability and clozapine gene-sets across more stringent thresholds of differentially expressed genes (i.e., <FDR 5%, <FDR 1%, and Bonferroni correction (Figure 4C). Such trends were not observed for HDL or any of the other traits tested, including SCZ (Figure 4C).

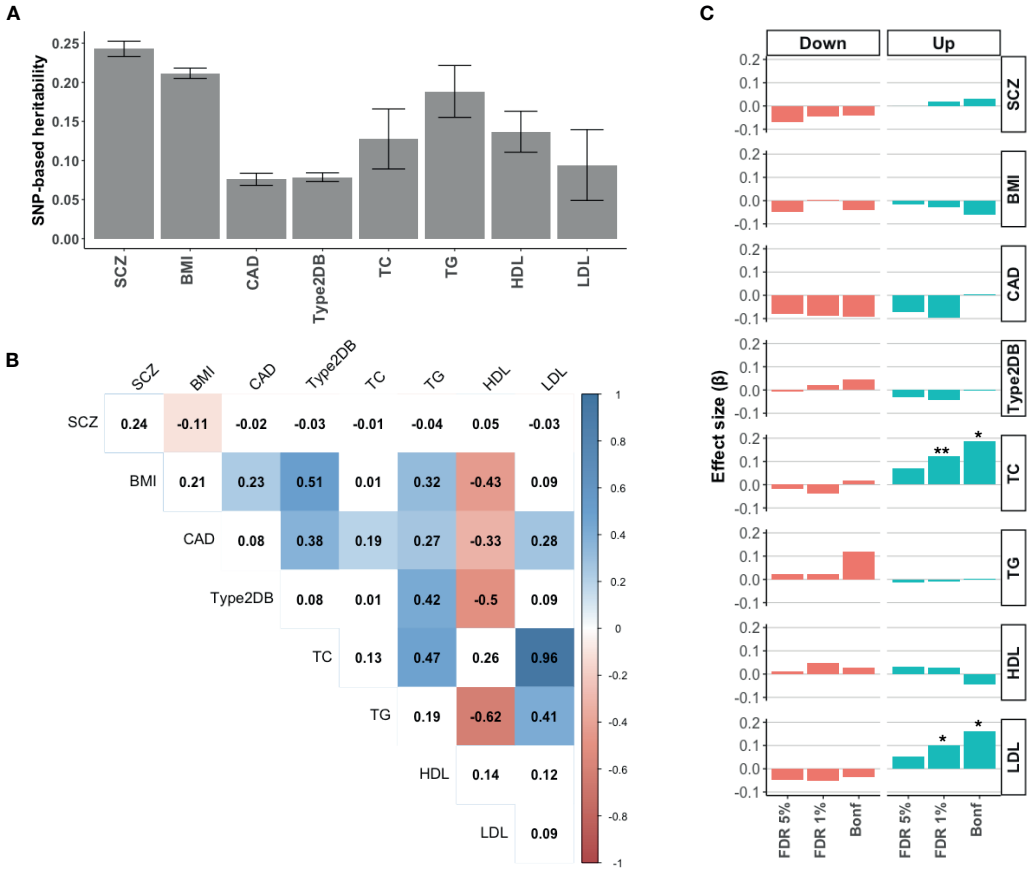


Figure 4. Clozapine gene-set analysis across cardiovascular traits. (A) SNP-based heritability estimates on the observed scale. (B) Genetic correlations between traits used in the analysis. The magnitude of the genetic correlation is only color-coded for estimates with p-value < 0.01. (C) MAGMA effect sizes (β) of GWAS trait associations with clozapine target genes are shown for down- and unregulated genes identified to be differentially expressed at FDR

*5%, FDR 1%, and Bonferroni correction thresholds. Asterisks denote nominal significance of *P<0.05, **P<0.01. SCZ = schizophrenia, BMI = body mass index, CAD = coronary artery disease, Type2DB = type 2 diabetes, TC = total cholesterol, TG = triglycerides, HDL = high-density-lipoprotein, LDL = low-density-lipoprotein.*

Discussion

We used an *in vitro* cell system of LCLs to study molecular effects of clozapine exposure to better understand the mechanisms involved in adverse effects of psychotropic drug use and make several important observations. First, genome-wide gene expression profiling of cells exposed to clozapine showed strong activation of cholesterol metabolism and deactivation of genes involved in cell cycle processes. These findings align with our observation of changed levels of DNAm upstream of the LDL-R and CCNF gene after clozapine exposure. Second, DNAm analyses suggest that effects of clozapine are time-dependent, indicating that time of drug exposure is an important experimental variable to take into account when studying clozapine. Third, integration of *in vitro* transcriptomic signatures with human tissues in GTEx highlight liver tissue and several immune and endocrine tissues as possible downstream effectors of clozapine exposure. Finally, genetic analysis of the results suggests that clozapine-response in LCLs is independent from schizophrenia disease risk and depleted from known drug targets of antipsychotic drugs, while it is likely linked to the genetic architecture involved in cholesterol and low-density lipoprotein levels.

in vitro gene expression changes after exposure to clozapine have been reported before for various cell lines. Here, we for the first time applied a genome-wide analysis of gene expression in lymphoblastoid cell lines in response to clozapine exposure. The central role of cholesterol metabolism in the response to clozapine is concordant with previously reported studies performed in other cell types and for different antipsychotic drugs(Ferno et al. 2011; Foley and Mackinnon 2014). Even though most of these were based on candidate genes rather than by genome-wide analyses, the results consistently implicate cholesterol metabolism in response to antipsychotic drugs (Raeder et al. 2006; Choi et al. 2009; Vik-Mo et al. 2009; Lauressergues et al. 2010, 2011; L.-H. Yang et al. 2007; Zhi Yang et al. 2009; Hu, Kutscher, and Davies 2010; J. Fernø et al. 2005; Liu et al. 2009; Lauressergues et al. 2012; Johan Fernø et al. 2009). At the gene network level, clozapine-associated genes group to clear gene clusters. In addition to cholesterol metabolism, we observed mitochondrial and nucleosome pathways to be upregulated after clozapine exposure, while cell cycle processes were downregulated, for example.

A small number of studies have identified a link between genetic markers in genes involved in cholesterol metabolism and adverse effects of antipsychotic treatment, helping to further substantiate these findings (Chowdhury, Remington, and Kennedy 2011; L. Yang et al. 2015, 2016). Additional studies suggest that alterations in cholesterol metabolism by antipsychotic drugs may contribute to the beneficial effects in treating psychosis. Cholesterol is extremely important in brain development and in sustaining neuronal connections and myelination (Dietschy 2009). Furthermore, Le Hellard et al. described an association between genes important in cholesterol metabolism (SREBP1 and SREBP2) and schizophrenia (Dietschy 2009; Le Hellard et al. 2010), which was confirmed in a schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014).

In addition to significantly upregulated genes, we also identified 284 genes that are downregulated in the presence of clozapine, with significant enrichment for genes involved in cell cycle pathways. A small number of studies have shown differing gene expression levels in cell cycle genes in patients with schizophrenia, compared to healthy controls (Z. Yang et al. 2017; Lin et al. 2016; Okazaki et al. 2016; Wang et al. 2010; Gassó et al. 2017). A similar connection with antipsychotic drug effects has not been reported before. It is possible that genes involved in cell cycle play a role in the etiology of schizophrenia, making it a possible target of antipsychotic drugs such as clozapine. Conversely, the described effect may be a direct consequence of clozapine toxicity. We have shown before that clozapine, in high concentrations, has a direct effect on viability of lymphoblastoid cell lines (de With et al. 2015). Downregulation of genes involved in cell cycle processes could thus be a direct (toxic) effect of clozapine. The toxic effects of clozapine and its metabolites have been associated with clozapine-induced agranulocytosis (de With et al. 2015; Williams et al. 1997; Pereira and Dean 2006; Lahdelma et al. 2010). We note that for this study, we used LCLs, which are blood-derived cells, but from a different progenitor cell and with different cellular functions than neutrophils. The question remains whether findings from *in vitro* LCL can be directly extrapolated to *in vivo* effects of clozapine on neutrophils.

While we found strong effects of clozapine exposure at the gene expression level, we did not observe similar global effects at the level of DNA methylation. We did find 3 differently methylated CpGs located within genes implicated by our gene expression analysis, including the LDL receptor gene. These effects, however, were not observed after 96h of clozapine exposure. Additionally, we observed overall enrichment of regulatory and immune pathways in the top associated DNAm probes. Although previous studies have indicated that antipsychotic drugs may induce changes in DNA methylation (Kinoshita et al. 2017; Burghardt et al. 2016; Houtepen et al. 2016; Rukova et al. 2014), we did not find evidence for immediate large effects on DNA methylation based on CpG sites assayed in our experiments. Possibly, subtle changes in DNA methylation patterns play a regulatory role in the observed gene expression changes but a larger sample size is needed to decipher these changes(Jones 2012). In addition, inclusion of more concentrations to which cells are exposed *in vitro* may provide finer experimental resolution to observe subtle epigenetic changes and identify key regulatory drivers. DNA methylation is one type of epigenetic regulation and assaying other regulatory mechanisms, such as histone modification or RNA regulation, may provide further insights into the regulatory dynamics that drive widespread changes in gene expression (Allis and Jenuwein 2016). Lastly, we observed DNA methylation changes to be time-dependent. While the clinical meaning of this remains speculative, our findings do suggest that the existing literature should be evaluated in the context of the duration of drug exposure. In addition, future studies could gain more insights into the function of clozapine when modeling drug exposure time as a variable in their analyses.

To further explore our gene expression findings, we set out to investigate these effects and their association with genetic susceptibility of schizophrenia. We did not observe significant enrichment of schizophrenia heritability across clozapine-associated genes. Two previous studies did report an association between schizophrenia heritability and antipsychotic target genes (Gaspar and Breen 2017; Skene et al. 2018). Their approach, however, differed from ours. While they used lists of antipsychotic target genes originating from pharmacological databases, we

used a list of experimentally derived gene sets after *in vitro* exposure to clozapine in LCLs. The genes we found to be differentially expressed did not overlap with antipsychotic drug target genes used in these two previous studies, which may explain the discrepancy in findings. If we assume that the clozapine-induced gene expression differences are a proxy for adverse effects, the lack of evidence of the enrichment analysis suggests that the genetic architecture of disease susceptibility is likely independent from susceptibility to adverse effects. We explored this further by examining the enrichment of genetic signal linked to cholesterol and cardiovascular disease-related traits such as body mass index, coronary artery disease, type 2 diabetes, and triglyceride levels. We observed a nominally significant increased enrichment for total cholesterol and LDL genetic signal in upregulated genes in response to clozapine exposure. While these findings did not survive multiple testing correction and should be interpreted with caution, the observed cholesterol and LDL heritability enrichment increased as we narrowed down on the most strongly associated clozapine genes suggesting that this association warrants further investigation. Mapping population-based heritability to *in vitro* experimental systems can then serve as a powerful approach to study biological pathways through integration of polygenic disease risk (Ori et al. 2019). Metabolic adverse effects are often observed in patients using clozapine (and antipsychotics in general) and improving our understanding of the mechanisms that underlie these adverse effects is an imperative area of research.

An important advantage of *in vitro* experimental models is the controlled laboratory environment, which not only decreases the signal-to-noise ratio in the collected data but also allows for precise manipulation of the model in follow up work. While our current results are consistent with previous findings in other cell types (Raeder et al. 2006; Choi et al. 2009; Vik-Mo et al. 2009; Lauressergues et al. 2010, 2011; L.-H. Yang et al. 2007; Zhi Yang et al. 2009; Hu, Kutscher, and Davies 2010; J. Fernø et al. 2005; Liu et al. 2009; Lauressergues et al. 2012; Johan Fernø et al. 2009), it remains unclear whether LCLs are an appropriate cell type to capture the molecular changes that are most relevant for studying adverse effects of antipsychotics in patients. To gain insight into the transferability of the *in vitro* clozapine-response in LCLs to functions of human tissues, we investigated how clozapine-associated genes are expressed across GTEx tissues. As gene expression patterns have been shown to have a significant degree of sharing across human tissues (GTEx Consortium et al. 2017; Buil et al., n.d.) effects discovered in one tissue may thus be informative for other tissues. We observed that preferentially expressed genes in GTEx-LCL tissue overlap the most with clozapine genes identified in LCLs *in vitro*, with whole blood ranked as second highest tissue. Preferentially expressed genes, however, represent only a subset of the clozapine genes. Using all associated genes, we demonstrated that upregulated clozapine genes have significantly different average expression in liver, muscle, lung, and testis tissue. Given the numerous and diverse set of reported possible adverse effects by clozapine treatment (Iqbal et al. 2003), it may not be unsurprising to find significant differences in mean expression of clozapine genes across multiple human tissues. Our finding of higher expression in the liver fits well the clinical presentation of clozapine treatment and the observed cholesterol gene ontology signature of upregulated clozapine genes in our assay. The liver is a central player in the function of cholesterol in the body and clozapine-related hepatotoxicity has furthermore been reported in cases of treatment with the drug (Keane et al. 2009; Kellner et al. 1993). We observe no differences for any of the GTEx brain tissues.

On top of within tissue effects, we also performed a more explorative analysis to examine clozapine gene expression-based similarity between tissues and found specific tissue pairs that were significantly different than tissue similarities based on expression of all genes. Here, we highlight some interesting observations. We observed a decrease in similarity for tissues involved in immune (spleen), endocrine (testis, ovary, adrenal and thyroid gland) and metabolic (adipose) functioning. The spleen and testis in particular stand out with a significant change in dissimilarity to many tissues. The spleen is the largest secondary lymphoid organ in the body and hosts a wide range of immunological functions, including the storage of leukocytes (Lewis, Williams, and Eisenbarth 2019). While it remains speculative if splenic dysfunction could for example lead to agranulocytosis, our findings do solicit for more research on the mechanism between clozapine and splenic function. An intriguing and perhaps also surprising finding is the implication of the testis among analysis of genes downregulated by clozapine. The testis has two primary functions; to produce sperm and to produce hormones, in particular testosterone, which is a sex steroid synthesized from cholesterol (Eacker et al. 2008). Genes downregulated by clozapine are enriched for cell division processes and cell proliferation is an important function of testicular cells (Sohni et al. 2019). While little is currently known about how clozapine or other antipsychotics may affect testicular function in adult humans, our finding does again point to cholesterol-related biology. In addition to the testis, we also observe deviation in between-tissue similarity for several other endocrine tissues. As endocrine (and metabolic) abnormalities are known causes of human obesity, the identified transcriptomic profile of clozapine and tissue relations may point to new avenues to study clozapine-induced weight gain (Baptista 1999). Together, these findings suggest that clozapine disproportionately affects the function of specific tissues. Our results furthermore demonstrate how experimental molecular signatures can be integrated with external genomic datasets, such as the GTEx project, to help translate *in vitro* findings to human biology.

While our work highlights the value of LCLs as an experimental tool to study the molecular mechanisms of clozapine response, particular in relation to possible molecular adverse effects, it does come with several limitations. First, we exposed the cells to supratherapeutic concentrations of clozapine to induce strong downstream molecular effects. In our gene expression analysis, we robustly identify hundreds of genes that showed a dose-response change in expression level. While these genes overlap with functional pathways that have previously been associated with clozapine as well, caution is still warranted in extrapolating these findings for clinical interpretation. Our work presents a first step in using LCLs as a model system but more research is needed to determine how these findings translate to *in vivo* molecular signatures observed in patients that are exposed to clinical concentrations of the drug. Second, our gene expression and DNAm experiments were conducted as separate experiments that used different concentrations and cell lines, which may have impacted our results. Future work should synchronize these experiments as much as possible. Third, we performed our experiment in a relatively small number of cell lines that, in the case of our DNAm experiment, were also derived from two nuclear families. Heterogeneity in clozapine response between cell lines likely exists and may have biased the findings of our study. Future work could use LCLs of different individuals than used in our study or use a larger number of cell lines that allows for correction of relatedness and other possible confounders.

We used LCLs as an *in vitro* model for studying molecular effects of clozapine exposure in an effort to improve our understanding of antipsychotic-induced adverse effects. Genome-wide gene expression profiling demonstrated a robust up-regulation of cholesterol metabolism and down-regulation of cell cycle pathways, with only limited changes in DNA methylation profiles. We did not find evidence that genes up- or down-regulated during clozapine exposure were enriched for genetic variation associated with schizophrenia. On the other hand, the observed enrichment signal with the genetic basis of total cholesterol and LDL levels and multi-tissue involved across immune, endocrine, and metabolic functions may provide important leads linked to antipsychotic drug induced metabolic adverse effects. The necessary challenge of large-scale, systematic prospective patient cohort studies of adverse effects of clozapine and other AP remains, while *in vitro* studies such as ours provide only glimpses of what may be relevant.

Declarations

Acknowledgements

The authors would like to thank Drs. Barbara Franke, Jeffrey Glennon, Jan Buitelaar, and Wouter Staal for their input in the early stages of the project and the GTEx Project and all used GWAS studies for making their data publicly available.

Author contributions

APSO and SAJW performed primary drafting of the manuscript with important input from RAO. SAJW, ES, and TW performed the experimental work. SAJW and APSO performed data analyses with input from SP. JV and JM helped process DNA methylation data and provided input on analyses. SJ and RAO oversaw the work. All authors reviewed and approved of the manuscript.

Funding

This work was partly supported by funding from the European Community's Seventh Framework Programme (FP7/2007-2013) Pediatric European Risperidone Studies (PERS) under grant agreement n°241959. RAO is supported by NIH grants R01 DA028526, MH090553, and MH078075. SDW was supported by a personal grant from Stichting de Drie Lichten. SDJ was part funded by NARSAD Young Investigator Grant (YI 60373). The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Competing interest

The authors declare no conflict of interest.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74.

Adkins, D. E., K. Aberg, J. L. McClay, J. Bukszár, Z. Zhao, P. Jia, T. S. Stroup, et al. 2011. "Genomewide Pharmacogenomic Study of Metabolic Side Effects to Antipsychotic Drugs." Molecular Psychiatry 16 (3): 321–32.

Allis, C. David, and Thomas Jenuwein. 2016. "The Molecular Hallmarks of Epigenetic Control." Nature Reviews. Genetics 17 (8): 487–500.

Andersohn, Frank, Christine Konzen, and Edeltraut Garbe. 2007. "Systematic Review: Agranulocytosis Induced by Nonchemotherapy Drugs." Annals of Internal Medicine 146 (9): 657–65.

Baptista, T. 1999. "Body Weight Gain Induced by Antipsychotic Drugs: Mechanisms and Management." Acta Psychiatrica Scandinavica 100 (1): 3–16.

Baumann, P., C. Hiemke, S. Ulrich, G. Eckermann, I. Gaertner, M. Gerlach, H-J Kuss, et al. 2004. "The AGNP-TDM Expert Group Consensus Guidelines: Therapeutic Drug Monitoring in Psychiatry." Pharmacopsychiatry 37 (6): 243–65.

Buil, Alfonso, Ana Viñuela, Andrew A. Brown, Matthew N. Davies, Ismael Padioleau, Deborah Bielser, Luciana Romano, et al. n.d. "Quantifying the Degree of Sharing of Genetic and Non-Genetic Causes of Gene Expression Variability across Four Tissues." https://doi.org/10.1101/053355.

Bulik-Sullivan, Brendan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Reprogen Consortium, Psychiatric Genomics Consortium, et al. 2015. "An Atlas of Genetic Correlations across Human Diseases and Traits." Nature Genetics 47 (11): 1237–41.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of The Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." Nature Genetics 47 (3): 291–95.

Burghardt, Kyle J., Jacyln M. Goodrich, Dana C. Dolinoy, and Vicki L. Ellingrod. 2016. "Gene-Specific DNA Methylation May Mediate Atypical Antipsychotic-Induced Insulin Resistance." Bipolar Disorders 18 (5): 423–32

Chen, Yi-An, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. 2013. "Discovery of Cross-Reactive Probes and Polymorphic CpGs in the Illumina Infinium HumanMethylation450 Microarray." Epigenetics: Official Journal of the DNA Methylation Society 8 (2): 203–9.

Choi, Kwang H., Brandon W. Higgs, Serge Weis, Jonathan Song, Ida C. Llenos, Jeannette R. Dulay, Robert H. Yolken, and Maree J. Webster. 2009. "Effects of Typical and Atypical Antipsychotic Drugs on Gene Expression Profiles in the Liver of Schizophrenia Subjects." BMC Psychiatry 9 (September): 57.

Chowdhury, Nabilah I., Gary Remington, and James L. Kennedy. 2011. "Genetics of Antipsychotic-Induced Side Effects and Agranulocytosis." Current Psychiatry Reports 13 (2): 156–65.

Cohen, D. 2014. "Prescribers Fear as a Major Side-Effect of Clozapine." Acta Psychiatrica Scandinavica.

Consortium, †the International Hapmap, and †The International HapMap Consortium. 2003. "The International HapMap Project." Nature.

Dietschy, John M. 2009. "Central Nervous System: Cholesterol Turnover, Brain Development and Neurodegeneration." Biological Chemistry.

Du, Pan, Warren a. Kibbe, and Simon M. Lin. 2008. "Lumi: A Pipeline for Processing Illumina Microarray." Bioinformatics 24 (13): 1547–48.

Du, Pan, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A. Kibbe, Lifang Hou, and Simon M. Lin. 2010. "Comparison of Beta-Value and M-Value Methods for Quantifying Methylation Levels by Microarray Analysis." BMC Bioinformatics 11 (1): 587.

Eacker, Stephen M., Nalini Agrawal, Kun Qian, Helén L. Dichek, Eun-Yeung Gong, Keesook Lee, and Robert E. Braun. 2008. "Hormonal Regulation of Testicular Steroid and Cholesterol Homeostasis." Molecular Endocrinology 22 (3): 623–35.

Ferno, Johan, Silje Skrede, Audun Osland Vik-Mo, Goran Jassim, Stephanie Le Hellard, and Vidar Martin Steen. 2011. "Lipogenic Effects of Psychotropic Drugs: Focus on the SREBP System." Frontiers in Bioscience 16 (January): 49–60.

Fernø, Johan, Audun O. Vik-Mo, Goran Jassim, Bjarte Håvik, Kjetil Berge, Silje Skrede, Oddrun A. Gudbrandsen, et al. 2009. "Acute Clozapine Exposure *in Vivo* Induces Lipid Accumulation and Marked Sequential Changes in the Expression of SREBP, PPAR, and LXR Target Genes in Rat Liver." Psychopharmacology 203 (1): 73–84.

Fernø, J., M. B. Raeder, A. O. Vik-Mo, S. Skrede, M. Glambek, K-J Tronstad, H. Breilid, et al. 2005. "Antipsychotic Drugs Activate SREBP-Regulated Expression of Lipid Biosynthetic Genes in Cultured Human Glioma Cells: A Novel Mechanism of Action?" The Pharmacogenomics Journal 5 (5): 298–304.

Foley, D. L., and A. Mackinnon. 2014. "A Systematic Review of Antipsychotic Drug Effects on Human Gene Expression Related to Risk Factors for Cardiovascular Disease." The Pharmacogenomics Journal 14 (5): 446–51.

Gaspar, H. A., and G. Breen. 2017. "Drug Enrichment and Discovery from Schizophrenia Genome-Wide Association Results: An Analysis and Visualisation Approach." Scientific Reports 7 (1): 12460.

Gassó, Patricia, Sergi Mas, Natalia Rodríguez, Daniel Boloc, Susana García-Cerro, Miquel Bernardo, Amalia Lafuente, and Eduard Parellada. 2017. "Microarray Gene-Expression Study in Fibroblast and Lymphoblastoid Cell Lines from Antipsychotic-Naïve First-Episode Schizophrenia Patients." Journal of Psychiatric Research 95 (December): 91–101.

Gebhardt, S., F. M. Theisen, M. Haberhausen, M. Heinzel-Gutenbrunner, P. M. Wehmeier, J. -C. Krieg, W. Kühnau, J. Schmidtke, H. Remschmidt, and J. Hebebrand. 2010. "Body Weight Gain Induced by Atypical Antipsychotics: An Extension of the Monocygotic Twin and Sib Pair Study." Journal of Clinical Pharmacy and Therapeutics.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." Nature 550 (7675): 204–13.

Harrison, Rebecca N. S., Robin M. Murray, Sang Hyuck Lee, Jose Paya Cano, David Dempster, Charles J. Curtis, Danai Dima, Fiona Gaughran, Gerome Breen, and Simone de Jong. 2016. "Gene-Expression Analysis of Clozapine Treatment in Whole Blood of Patients with Psychosis." Psychiatric Genetics 26 (5): 211–17.

Horvath, Steve. 2011. Weighted Network Analysis: Applications in Genomics and Systems Biology. Springer Science & Business Media.

Houtepen, Lotte C., Annet H. van Bergen, Christiaan H. Vinkers, and Marco P. M. Boks. 2016. "DNA Methylation Signatures of Mood Stabilizers and Antipsychotics in Bipolar Disorder." Epigenomics 8 (2): 197–208.

Huang, Da Wei, Richard a. Lempicki, and Brad T. Sherman. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." Nature Protocols 4 (1): 44–57.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." Nucleic Acids Research 37 (1): 1–13.

Hu, Yueshan, Eric Kutscher, and Gareth E. Davies. 2010. "Berberine Inhibits SREBP-1-Related Clozapine and Risperidone Induced Adipogenesis in 3T3-L1 Cells." Phytotherapy Research: PTR 24 (12): 1831–38.

Iqbal, Mohammad Masud, Atiq Rahman, Zahid Husain, Syed Zaber Mahmud, Wlliam G. Ryan, and Jacqueline M. Feldman. 2003. "Clozapine: A Clinical Review of Adverse Effects and Management." Annals of Clinical Psychiatry: Official Journal of the American Academy of Clinical Psychiatrists 15 (1): 33–48.

Jones, Peter A. 2012. "Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and beyond." Nature Reviews. Genetics 13 (7): 484–92.

Kane, J., G. Honigfeld, J. Singer, and H. Meltzer. 1988. "Clozapine for the Treatment-Resistant Schizophrenic. A Double-Blind Comparison with Chlorpromazine." Archives of General Psychiatry 45 (9): 789–96.

Keane, Sarah, Abbie Lane, Terence Larkin, and Mary Clarke. 2009. "Management of Clozapine-Related Hepatotoxicity." Journal of Clinical Psychopharmacology 29 (6): 606–7.

Kellner, M., K. Wiedemann, J. C. Krieg, and P. A. Berg. 1993. "Toxic Hepatitis by Clozapine Treatment." The American Journal of Psychiatry 150 (6): 985–86.

Kinoshita, Makoto, Shusuke Numata, Atsushi Tajima, Hidenaga Yamamori, Yuka Yasuda, Michiko Fujimoto, Shinya Watanabe, et al. 2017. "Effect of Clozapine on DNA Methylation in Peripheral Leukocytes from Patients with Treatment-Resistant Schizophrenia." International Journal of Molecular Sciences 18 (3)

Lahdelma, Liisa, Sofia Oja, Matti Korhonen, and Leif C. Andersson. 2010. "Clozapine Is Cytotoxic to Primary Cultures of Human Bone Marrow Mesenchymal Stromal Cells." Journal of Clinical Psychopharmacology 30 (4): 461–63.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." BMC Bioinformatics 9: 559.

Lauressergues, Emilie, Elodie Bert, Patrick Duriez, Dean Hum, Zouher Majd, Bart Staels, and Didier Cussac. 2012. "Does Endoplasmic Reticulum Stress Participate in APD-Induced Hepatic Metabolic Dysregulation?" Neuropharmacology 62 (2): 784–96.

Lauressergues, Emilie, Françoise Martin, Audrey Helleboid, Emmanuel Bouchaert, Didier Cussac, Régis Bordet, Dean Hum, et al. 2011. "Overweight Induced by Chronic Risperidone Exposure Is Correlated with Overexpression of the SREBP-1c and FAS Genes in Mouse Liver." Naunyn-Schmiedeberg's Archives of Pharmacology 383 (4): 423–36.

Lauressergues, Emilie, Bart Staels, Karine Valeille, Zouher Majd, Dean W. Hum, Patrick Duriez, and Didier Cussac. 2010. "Antipsychotic Drug Action on SREBPs-Related Lipogenesis and Cholesterogenesis in Primary Rat Hepatocytes." Naunyn-Schmiedeberg's Archives of Pharmacology 381 (5): 427–39.

Leeuw, Christiaan A. de, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. 2015. "MAGMA: Generalized Gene-Set Analysis of GWAS Data." PLoS Computational Biology 11 (4).

Le Hellard, S., T. W. Mühleisen, S. Djurovic, J. Fernø, Z. Ouriaghi, M. Mattheisen, C. Vasilescu, et al. 2010. "Polymorphisms in SREBF1 and SREBF2, Two Antipsychotic-Activated Transcription Factors Controlling Cellular Lipogenesis, Are Associated with Schizophrenia in German and Scandinavian Samples." Molecular Psychiatry 15 (5): 463–72.

Lett, T. A. P., T. J. M. Wallace, N. I. Chowdhury, A. K. Tiwari, J. L. Kennedy, and D. J. Müller. 2012. "Pharmacogenetics of Antipsychotic-Induced Weight Gain: Review and Clinical Implications." Molecular Psychiatry 17 (3): 242–66.

Leucht, Stefan, Andrea Cipriani, Loukia Spineli, Dimitris Mavridis, Deniz Orey, Franziska Richter, Myrto Samara, et al. 2013. "Comparative Efficacy and Tolerability of 15 Antipsychotic Drugs in Schizophrenia: A Multiple-Treatments Meta-Analysis." The Lancet 382 (9896): 951– 62.

Lewis, Steven M., Adam Williams, and Stephanie C. Eisenbarth. 2019. "Structure and Function of the Immune System in the Spleen." Science Immunology 4 (33).

Leykin, I., R. Mayer, and M. Shinitzky. 1997. "Short-and Long-Term Immunosuppressive Effects of Clozapine and Haloperidol." Immunopharmacology.

Lin, Mingyan, Erika Pedrosa, Anastasia Hrabovsky, Jian Chen, Benjamin R. Puliafito, Stephanie R. Gilbert, Deyou Zheng, and Herbert M. Lachman. 2016. "Integrative Transcriptome Network Analysis of iPSC-Derived Neurons from Schizophrenia and Schizoaffective Disorder Patients with 22q11.2 Deletion." BMC Systems Biology 10 (1): 105.

Liu, Yanhong, Ronald Jandacek, Therese Rider, Patrick Tso, and Robert K. McNamara. 2009. "Elevated Delta-6 Desaturase (FADS2) Expression in the Postmortem Prefrontal Cortex of Schizophrenic Patients: Relationship with Fatty Acid Composition." Schizophrenia Research 109 (1-3): 113–20.

Malhotra, Anil K., Christoph U. Correll, Nabilah I. Chowdhury, Daniel J. Müller, Peter K. Gregersen, Annette T. Lee, Arun K. Tiwari, et al. 2012. "Association between Common Variants near the Melanocortin 4 Receptor Gene and Severe Antipsychotic Drug-Induced Weight Gain." Archives of General Psychiatry 69 (9): 904–12.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28 (5): 495–501.

Melas, Philippe A., Maria Rogdaki, Urban Ösby, Martin Schalling, Catharina Lavebratt, and Tomas J. Ekström. 2012. "Epigenetic Aberrations in Leukocytes of Patients with Schizophrenia: Association of Global DNA Methylation with Antipsychotic Drug Treatment and Disease Onset." FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology 26 (6): 2712–18.

Melé, Marta, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, et al. 2015. "Human Genomics. The Human Transcriptome across Tissues and Individuals." Science 348 (6235): 660–65.

Morag, Ayelet, Julia Kirchheiner, Moshe Rehavi, and David Gurwitz. 2010. "Human Lymphoblastoid Cell Line Panels: Novel Tools for Assessing Shared Drug Pathways." Pharmacogenomics 11 (3): 327–40.

Müller, Daniel J., Nabilah I. Chowdhury, and Clement C. Zai. 2013. "The Pharmacogenetics of Antipsychotic-Induced Adverse Events." Current Opinion in Psychiatry 26 (2): 144–50.

Nikpay, Majid, Anuj Goel, Hong-Hee Won, Leanne M. Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, et al. 2015. "A Comprehensive 1,000 Genomes-Based Genome-Wide Association Meta-Analysis of Coronary Artery Disease." Nature Genetics 47 (10): 1121– 30.

Okazaki, Satoshi, Shuken Boku, Ikuo Otsuka, Kentaro Mouri, Shinsuke Aoyama, Kyoichi Shiroiwa, Ichiro Sora, et al. 2016. "The Cell Cycle-Related Genes as Biomarkers for Schizophrenia." Progress in Neuro-Psychopharmacology & Biological Psychiatry 70 (October): 85–91.

Ori, Anil P. S., Merel H. M. Bot, Remco T. Molenhuis, Loes M. Olde Loohuis, and Roel A. Ophoff. 2019. "A Longitudinal Model of Human Neuronal Differentiation for Functional Investigation of Schizophrenia Polygenic Risk." Biological Psychiatry 85 (7): 544–53.

Ota, Vanessa Kiyomi, Cristiano Noto, Ary Gadelha, Marcos Leite Santoro, Leticia Maria Spindola, Eduardo Sauerbronn Gouvea, Roberta Sessa Stilhano, et al. 2014. "Changes in Gene Expression and Methylation in the Blood of Patients with First-Episode Psychosis." Schizophrenia Research 159 (2-3): 358–64.

Pereira, Avril, and Brian Dean. 2006. "Clozapine Bioactivation Induces Dose-Dependent, Drug-Specific Toxicity of Human Bone Marrow Stromal Cells: A Potential *in vitro* System for the Study of Agranulocytosis." Biochemical Pharmacology.

Pidsley, Ruth, Chloe C. Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C. Schalkwyk. 2013. "A Data-Driven Approach to Preprocessing Illumina 450K Methylation Array Data." BMC Genomics 14: 293.

Price, Magda E., Allison M. Cotton, Lucia L. Lam, Pau Farré, Eldon Emberly, Carolyn J. Brown, Wendy P. Robinson, and Michael S. Kobor. 2013. "Additional Annotation Enhances Potential for Biologically-Relevant Analysis of the Illumina Infinium HumanMethylation450 BeadChip Array." Epigenetics & Chromatin 6 (1): 4.

Raeder, Maria B., Johan Fernø, Audun O. Vik-Mo, and Vidar M. Steen. 2006. "SREBP Activation by Antipsychotic- and Antidepressant-Drugs in Cultured Human Liver Cells: Relevance for Metabolic Side-Effects?" Molecular and Cellular Biochemistry 289 (1-2): 167– 73.

Roerig, James L., Kristine J. Steffen, and James E. Mitchell. 2011. "Atypical Antipsychotic-Induced Weight Gain: Insights into Mechanisms of Action." CNS Drugs 25 (12): 1035–59.

Roth, Bryan L., Estelle Lopez, Shamil Patel, and Wesley K. Kroeze. 2000. "The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?" The Neuroscientist.

Rukova, Blaga, Rada Staneva, Savina Hadjidekova, Georgi Stamenov, Vihra Milanova, and Draga Toncheva. 2014. "Whole Genome Methylation Analyses of Schizophrenia Patients before and after Treatment." Biotechnology, Biotechnological Equipment 28 (3): 518–24.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." Nature.

Scott, Robert A., Laura J. Scott, Reedik Mägi, Letizia Marullo, Kyle J. Gaulton, Marika Kaakinen, Natalia Pervjakova, et al. 2017. "An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans." Diabetes 66 (11): 2888–2902.

Skene, Nathan G., Julien Bryois, Trygve E. Bakken, Gerome Breen, James J. Crowley, Héléna A. Gaspar, Paola Giusti-Rodriguez, et al. 2018. "Genetic Identification of Brain Cell Types Underlying Schizophrenia." Nature Genetics 50 (6): 825–33.

Smyth, G. K. n.d. "Limma: Linear Models for Microarray Data." Bioinformatics and Computational Biology Solutions Using R and Bioconductor.

Sohni, Abhishek, Kun Tan, Hye-Won Song, Dana Burow, Dirk G. de Rooij, Louise Laurent, Tung-Chin Hsieh, et al. 2019. "The Neonatal and Adult Human Testis Defined at the Single-Cell Level." Cell Reports 26 (6): 1501–17. e4.

Stapel, Britta, Alexandra Kotsiari, Michaela Scherr, Denise Hilfiker-Kleiner, Stefan Bleich, Helge Frieling, and Kai G. Kahl. 2017. "Olanzapine and Aripiprazole Differentially Affect Glucose Uptake and Energy Metabolism in Human Mononuclear Blood Cells." Journal of Psychiatric Research 88 (May): 18–27.

Stroup, T. Scott, Tobias Gerhard, Stephen Crystal, Cecilia Huang, and Mark Olfson. 2016. "Comparative Effectiveness of Clozapine and Standard Antipsychotic Treatment in Adults With Schizophrenia." The American Journal of Psychiatry 173 (2): 166–73.

Surakka, Ida, Momoko Horikoshi, Reedik Mägi, Antti-Pekka Sarin, Anubha Mahajan, Vasiliki Lagou, Letizia Marullo, et al. 2015. "The Impact of Low-Frequency and Rare Variants on Lipid Levels." Nature Genetics 47 (6): 589–97.

Swathy, Babu, and Moinak Banerjee. 2017. "Haloperidol Induces Pharmacoepigenetic Response by Modulating miRNA Expression, Global DNA Methylation and Expression Profiles of Methylation Maintenance Genes and Genes Involved in Neurotransmission in Neuronal Cells." PloS One 12 (9): e0184209.

Swathy, Babu, Koramannil R. Saradalekshmi, Indu V. Nair, Chandrasekharan Nair, and Moinak Banerjee. 2017. "Pharmacoepigenomic Responses of Antipsychotic Drugs on Pharmacogenes Are Likely to Be Modulated by miRNAs." Epigenomics 9 (6): 811–21.

Taylor, David M. 2017. "Clozapine for Treatment-Resistant Schizophrenia: Still the Gold Standard?" CNS Drugs 31 (3): 177–80.

Tschen, A. C., M. J. Rieder, L. K. Oyewumi, and D. J. Freeman. 1999. "The Cytotoxicity of Clozapine Metabolites: Implications for Predicting Clozapine-Induced Agranulocytosis." Clinical Pharmacology and Therapeutics 65 (5): 526–32.

Venables, W. N., and B. D. Ripley. 2002. "Modern Applied Statistics with S." Statistics and Computing.

Vik-Mo, Audun O., Astrid B. Birkenaes, Johan Fernø, Halldora Jonsdottir, Ole A. Andreassen, and Vidar M. Steen. 2008. "Increased Expression of Lipid Biosynthesis Genes in Peripheral Blood Cells of Olanzapine-Treated Patients." The International Journal of Neuropsychopharmacology / Official Scientific Journal of the Collegium Internationale Neuropsychopharmacologicum 11 (5): 679–84.

Vik-Mo, Audun O., Johan Fernø, Silje Skrede, and Vidar M. Steen. 2009. "Psychotropic Drugs up-Regulate the Expression of Cholesterol Transport Proteins Including ApoE in Cultured Human CNS- and Liver Cells." BMC Pharmacology 9 (August): 10.

Wagner, Alex H., Adam C. Coffman, Benjamin J. Ainscough, Nicholas C. Spies, Zachary L. Skidmore, Katie M. Campbell, Kilannin Krysiak, et al. 2016. "DGIdb 2.0: Mining Clinically Relevant Drug–gene Interactions." Nucleic Acids Research.

Wang, L., H. E. Lockstone, P. C. Guest, Y. Levin, A. Palotás, S. Pietsch, E. Schwarz, et al. 2010. "Expression Profiling of Fibroblasts Identifies Cell Cycle Abnormalities in Schizophrenia." Journal of Proteome Research 9 (1): 521–27.

Weiden, Peter J., Joan A. Mackell, and Diana D. McDonnell. 2004. "Obesity as a Risk Factor for Antipsychotic Noncompliance." Schizophrenia Research.

Welsh, Marleen, Lara Mangravite, Marisa Wong Medina, Kelan Tantisira, Wei Zhang, R. Stephanie Huang, Howard McLeod, and M. Eileen Dolan. 2009. "Pharmacogenomic Discovery Using Cell-Based Models." Pharmacological Reviews 61 (4): 413–29.

Wen, Yujia, Eric R. Gamazon, Wasim K. Bleibel, Claudia Wing, Shuangli Mi, Bridget E. McIlwee, Shannon M. Delaney, Shiwei Duan, Hae Kyung Im, and M. Eileen Dolan. 2012. "An eQTL-Based Method Identifies CTTN and ZMAT3 as Pemetrexed Susceptibility Markers." Human Molecular Genetics 21 (7): 1470–80.

Williams, D. P., M. Pirmohamed, D. J. Naisbitt, J. L. Maggs, and B. K. Park. 1997. "Neutrophil Cytotoxicity of the Chemically Reactive Metabolite(s) of Clozapine: Possible Role in Agranulocytosis." The Journal of Pharmacology and Experimental Therapeutics 283 (3): 1375–82.

With, S. A. J. de, S. L. Pulit, T. Wang, W. G. Staal, W. W. van Solinge, P. I. W. de Bakker, and R. A. Ophoff. 2015. "Genome-Wide Association Study of Lymphoblast Cell Viability after Clozapine Exposure." American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 168B (2): 116–22.

Yang, Li-Hung, Tzer-Ming Chen, Sung-Tsai Yu, and Yen-Hui Chen. 2007. "Olanzapine Induces SREBP-1-Related Adipogenesis in 3T3-L1 Cells." Pharmacological Research: The Official Journal of the Italian Pharmacological Society 56 (3): 202–8.

Yang, Lin, Jianhua Chen, Dengtang Liu, Shunying Yu, Enzhao Cong, Yan Li, Haisu Wu, et al. 2015. "Association between SREBF2 Gene Polymorphisms and Metabolic Syndrome in Clozapine-Treated Patients with Schizophrenia." Progress in Neuro-Psychopharmacology & Biological Psychiatry 56 (January): 136–41.

Yang, Lin, Jianhua Chen, Yan Li, Yan Wang, Shiqiao Liang, Yongyong Shi, Shenxun Shi, and Yifeng Xu. 2016. "Association between SCAP and SREBF1 Gene Polymorphisms and Metabolic Syndrome in Schizophrenia Patients Treated with Atypical Antipsychotics." The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry 17 (6): 467–74.

Yang, Zhi, Ji-Ye Yin, Zhi-Cheng Gong, Qiong Huang, Hao Chen, Wei Zhang, Hong-Hao Zhou, and Zhao-Qian Liu. 2009. "Evidence for an Effect of Clozapine on the Regulation of Fat-Cell Derived Factors." Clinica Chimica Acta; International Journal of Clinical Chemistry 408 (1-2): 98–104.

Yang, Z., M. Li, X. Hu, B. Xiang, W. Deng, Q. Wang, Y. Wang, et al. 2017. "Rare Damaging Variants in DNA Repair and Cell Cycle Pathways Are Associated with Hippocampal and Cognitive Dysfunction: A Combined Genetic Imaging Study in First-Episode Treatment-Naive Patients with Schizophrenia." Translational Psychiatry 7 (2): e1028.

Yan, Hu, Jin-Dong Chen, and Xiao-Yan Zheng. 2013. "Potential Mechanisms of Atypical Antipsychotic-Induced Hypertriglyceridemia." Psychopharmacology 229 (1): 1–7.

Yengo, Loic, Julia Sidorenko, Kathryn E. Kemper, Zhili Zheng, Andrew R. Wood, Michael N. Weedon, Timothy M. Frayling, et al. n.d. "Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ~ 700,000 Individuals of European Ancestry."

Zhang, Bin, and Steve Horvath. 2005. "A General Framework for Weighted Gene Co-Expression Network Analysis." Statistical Applications in Genetics and Molecular Biology 4 (August): Article17.

Chapter 4 - Supplemental Materials

Full supplemental materials can be found here:
https://www.biorxiv.org/content/10.1101/2020.09.22.308262v1.supplementary-material

S1. Supplementary Material and Methods

S1.1 Clozapine exposure and in vitro experimental design

We performed drug exposure experiments in 6-well plates (Genesee, San Diego, CA, USA) and assessed cell viability using the TC10 automated cell counter (Bio-Rad, Hercules, CA, USA), according to manufacturer's instructions. Clozapine was purchased from Sigma Aldrich (St. Louis, MO, USA). Previous work has suggested that clinical concentrations of antipsychotics may not induce significant gene expression changes *in vitro* (Fernø et al. 2006). There is evidence that *in vivo* concentrations of antipsychotic drugs are higher in brain tissue than in peripheral blood (Weigmann et al. 1999; Kornhuber et al. 1999). We therefore chose to expose cell lines to different clozapine concentrations, with clinical concentration set at 2μM (Baumann et al. 2004). Clozapine was dissolved in culture medium with dimethyl sulfoxide (DMSO), with a maximum concentration of 0.1%. Cell lines were exposed for 24 hours to clinical concentration, 10x, 50x and 100x clinical concentration (20μM-100 μM-200 μM clozapine) and vehicle (DMSO); each concentration was measured in 4 cell lines, after which RNA was obtained for gene expression analysis.

To study DNA methylation changes in response to clozapine, we performed an experiment similar to the gene expression study; LCLs were exposed to different concentrations of clozapine (vehicle (DMSO), 1x, 20x, 40x and 60 times clinical concentration) and exposure times were 24h and 96h (Supplemental Figure 1B).

S1.2. Sample processing and gene expression data

We performed RNA extraction with Qiagen RNeasy mini kit (Qiagen, Valencia, CA, USA), according to manufacturer's instructions. RNA quantity and quality were measured with T2100 BioAnalyzer (Agilent, Santa Clara, CA, USA) and verified with a NanoDrop Spectrophotometer (NanoDrop products, Wilmington, DE, USA). Gene expression profiling was carried out using Illumina® HumanHT-12 v4 Expression BeadChip technology (Illumina, San Diego, CA, USA).

S1.3. Sample processing and DNA methylation data

After desired exposure time, cells were lysed and DNA was extracted with DNeasy® Blood and Tissue kit (Qiagen, Valencia, CA, USA), according to the instructions of the manufacturer. DNA quality and quantity were assessed with the picogreen® assay (VWR, West Chester, PA, USA) and Nanodrop (ThermoScientific, Wilmington, DE, USA). DNA methylation assays were performed with Illumina® Infinium HumanMethylation450 Beadchip arrays (Illumina, San Diego, CA, USA), assaying approximately 450,000 CpG sites.

Gene ontology analysis

We performed functional gene ontology analysis using DAVID (Database for Annotation, Visualization and Integrated Discovery, version 6.8, interrogated February 2018) (Huang, Sherman, and Lempicki 2009; Huang, Lempicki, and Sherman 2009), with default settings.

S1.4. Functional enrichment analysis of DNA methylation data

Genomic Regions Enrichment of Annotations Tools (GREAT, v3.0) was used to predict the biological function of the top methylation probes associated to clozapine exposure. GREAT links both proximal and distal genomic CpG sites with their putative target genes and implements both a gene-based test and a region-based test using the hypergeometric and binomial test, respectively, to assess enrichment of genomic regions in biological annotations (McLean et al. 2010). CpG sites were uploaded to the GREAT web portal (http://great.stanford.edu/public/html/) and analyses were run using the hg19 reference annotation and the whole genome as background. Genomic regions were assigned to genes if they are between 5 Kb upstream and 1 Kb downstream of the TSS, plus up to 1 Mb distal. Pathway annotations from GO Biological Processes, GO Cellular Component, GO Molecular Function, MSigDB, and PANTHER were used to infer biological meaning for CpG sites associated with clozapine.

S2. Supplementary Results

S2.1. Clozapine-associated genes and their preferential expression in GTEx tissues

To investigate tissue-specificity of genes identified to be differentially expressed after clozapine exposure, we used an available list of genes that are preferentially expressed in an individual tissue as identified by GTEx (Melé et al. 2015). Preferential tissue expression is defined as all instances where the mean expression of the gene in the tested tissue was significantly higher (FDR<0.01 and a log2 fold change >= 4) than in the samples from the rest of the tissues. Visualized below is the fraction of genes with tissue preferential expression that are detected in our assay (x-axis) versus detected in our assay and associated to clozapine exposure (FDR < 5%) for each tissue. We observe that almost half of the differentially expressed genes have preferential expression in LCL tissue in GTEx. The second highest tissue is whole blood.

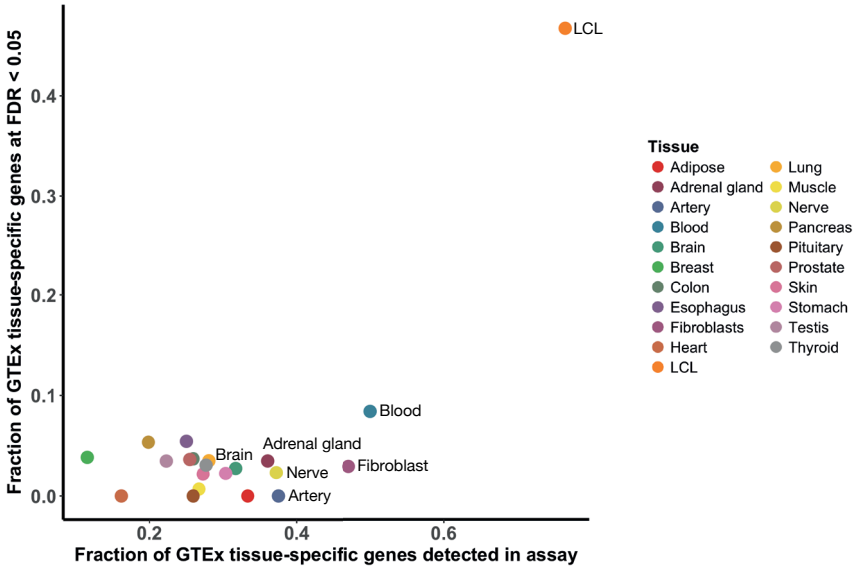
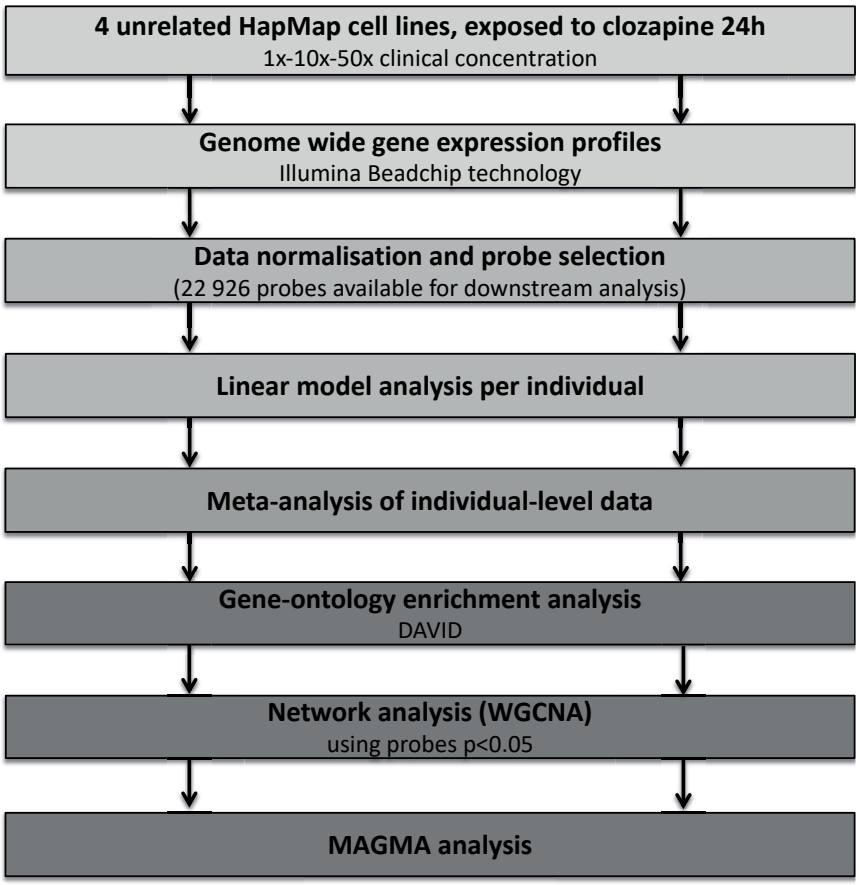
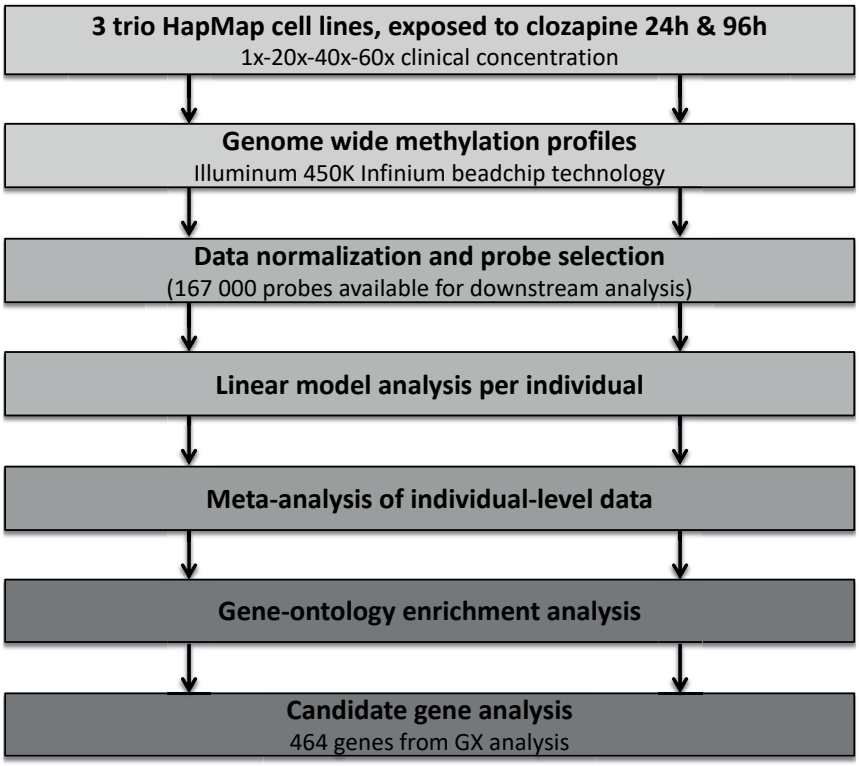


Figure S2. Differentially expressed genes are preferentially expressed in LCL tissue in GTEx.

Supplementary Figures



Supplementary Figure 1A. Flow chart of gene expression analyses plan.



Supplementary Figure 1B. Flow chart of DNA methylation analyses plan.

Supplementary Table 1. List of differentially expressed genes

Available as excel sheet online.

24h clozapine exposure					96h clozapine exposure				
Probe ID	Chr#	Coordinate	Annotation	P-value*	Probe ID	Chr#	Coordinate	Annotation	P-value*
cg09495017	16	57,124,763	*CNOT1*	$5.40^{*}10^{-7}$	cg21293934	18	14,738,230	*ANKRD30B*	$8.87^{*}10^{-7}$
cg16258062	2	234,048,887	-	$5.60^{*}10^{-7}$	cg19182557	2	130,061,825	-	$9.21^{*}10^{-7}$
cg16258062	1	47,654,324	*FOXE3*	$1.15^{*}10^{-6}$	cg15463280	11	95,955,596	-	$2.36^{*}10^{-6}$
cg15066636	6	33,187,127	*HLA-DPB2*	$1.42^{*}10^{-6}$	cg12564567	11	115,876,398	-	$5.87^{*}10^{-6}$
cg17488052	1	77,993,925	*USP33*	$1.48^{*}10^{-6}$	cg26647200	16	2,422,776	*CCNF*	$6.87^{*}10^{-6}$
cg25181236	4	56,082,032	*CLOCK*	$1.48^{*}10^{-6}$	cg09333631	3	44,777,608	*KIF15, KIAA1143*	$7.99^{*}10^{-6}$
cg01531409	14	59,781,022	*PPM1A*	$2.07^{*}10^{-6}$	cg16924010	3	195,500,852	-	$8.51^{*}10^{-6}$
cg09840472	7	22,730,922	-	$2.27^{*}10^{-6}$	cg01842314	10	106,102,325	*CCDC147*	$1.07^{*}10^{-5}$
cg24207009	17	73,549,157	*TNRC6C*	$2.37^{*}10^{-6}$	cg23898204	2	724,927	-	$1.11^{*}10^{-5}$
cg27170003	17	3,713,677	*CAMKK1*	$2.39^{*}10^{-6}$	cg15000279	19	33,976,849	-	$1.28^{*}10^{-5}$

Supplementary Table 2. Top 10 DNA methylation probes after 24h and 96h of clozapine exposure.

24h clozapine exposure				
Probe ID	Annotation	P-value	Gene expression Probe ID	Gene expression P-value
cg05455234	*PCNT (Pericentrin)*	$1.98^{*}10^{-5}$	ILMN_1810922	$5.68^{*}10^{-8}$
cg22971501	*LDLR (Low Density Lipoprotein Receptor)*	$4.75^{*}10^{-5}$	ILMN_2053415	$3.97^{*}10^{-14}$
cg01233620	*CLEC16A (C-Type Lectin Domain Containing 16A)*	$6.30^{*}10^{-5}$	ILMN_1781752	$1.04^{*}10^{-7}$
96h clozapine exposure				
cg26647200	*CCNF (Cyclin F)*	$6.87^{*}10^{-6}$	ILMN_1773119	$4.41^{*}10^{-11}$

Supplementary Table 3. Candidate gene analysis: significant methylation probes after 24h and 96h of clozapine exposure.

Set	Bonferroni	FDR < 1% (q<0.01)	FDR<5% (q<0.05)
Differentially expressed genes	p = 0.92 (311 genes)	p = 0.74 (919 genes)	p = 0.22 (1543 genes)
Upregulated genes	p = 0.75 (138 genes)	p = 0.71 (457 genes)	p = 0.99 (772 genes)
Downregulated genes	p = 0.64 (173 genes)	p = 0.39 (462 genes)	p = 0.09 (771 genes)

Supplementary Table 8. Gene-set analyses results of the schizophrenia GWAS at varying significance levels.

References – supplementary information

Baumann, P., C. Hiemke, S. Ulrich, G. Eckermann, I. Gaertner, M. Gerlach, H-J Kuss, et al. 2004. "The AGNP-TDM Expert Group Consensus Guidelines: Therapeutic Drug Monitoring in Psychiatry." Pharmacopsychiatry 37 (6): 243–65.

Fernø, Johan, Silje Skrede, Audun O. Vik-Mo, Bjarte Håvik, and Vidar M. Steen. 2006. "Drug- Induced Activation of SREBP-Controlled Lipogenic Gene Expression in CNS-Related Cell Lines: Marked Differences between Various Antipsychotic Drugs." BMC Neuroscience 7 (October): 69.

Huang, Da Wei, Richard a. Lempicki, and Brad T. Sherman. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." Nature Protocols 4 (1): 44–57.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." Nucleic Acids Research 37 (1): 1–13.

Kornhuber, J., A. Schultz, J. Wiltfang, I. Meineke, C. H. Gleiter, R. Zöchling, K. W. Boissl, F. Leblhuber, and P. Riederer. 1999. "Persistence of Haloperidol in Human Brain Tissue." The American Journal of Psychiatry 156 (6): 885–90.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28 (5): 495–501.

Melé, Marta, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, et al. 2015. "Human Genomics. The Human Transcriptome across Tissues and Individuals." Science 348 (6235): 660–65.

Weigmann, H., S. Härtter, V. Fischer, N. Dahmen, and C. Hiemke. 1999. "Distribution of Clozapine and Desmethylclozapine between Blood and Brain in Rats." European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology 9 (3): 253–56.

PART 2

DNA methylation algorithms and the role of biological age in schizophrenia

CHAPTER 5

A systematic evaluation of 41 DNA methylation predictors across 101 data preprocessing and normalization strategies highlights considerable variation in algorithm performance

Authors
Anil P.S. Ori[1]
Ake T Lu[2]
Steve Horvath[2,3]
Roel A Ophoff[1,3,4]

Affiliations
[1.] University of California Los Angeles, Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA
[2.] University of California Los Angeles, Department of Human Genetics, David Geffen, School of Medicine , Los Angeles, CA, USA
[3.] University of California Los Angeles, Department of Biostatistics, Fielding School of Public Health, Los Angeles, CA, USA
[4.] Erasmus University Medical Center, Department of Psychiatry, Rotterdam, The Netherlands

Abstract

DNA methylation (DNAm) based predictors hold great promise to serve as clinical tools for health interventions and disease management. While these algorithms often have high prediction accuracy and are associated with many disease-related phenotypes, the reliability of their performance remains to be determined. We therefore conducted a systematic evaluation across 101 different data processing strategies that preprocess and normalize DNAm data and assessed how each analytical strategy affects the reliability and prediction accuracy of 41 DNAm-based predictors. Our analyses were conducted in a large EPIC DNAm sample of the Jackson Heart Study (N=2,053) that included 146 pairs of technical replicate samples. By estimating the average absolute agreement between replicate pairs, we show that 32 out of 41 predictors (78%) demonstrate excellent test-retest reliability when appropriate data processing and normalization steps are implemented. Across all pairs of predictors, we find a moderate correlation in performance across analytical strategies (mean rho=0.40, SD=0.27), highlighting significant heterogeneity in performance across algorithms within a choice of an analytical pipeline. (Un)successful removal of technical variation furthermore significantly impacts downstream phenotypic association analysis, such as all-cause mortality risk associations. We show that DNAm-based algorithms are sensitive to technical variation. The right choice of data processing and normalization pipeline is important to achieve reproducible estimates and improve prediction accuracy in downstream phenotypic association analyses. For each of the 41 DNAm predictors, we report its test-retest reliability and provide the best performing analytical strategy as a guideline for the research community. As DNAm-based predictors become more and more widely used, both for research purposes School of Medicine as well as for clinic applications, our work helps improve their performance and standardize their implementation.

Manuscript status: in submission
Preprint available: https://www.biorxiv.org/content/10.1101/2021.09.29.462387v1

Introduction

DNA methylation (DNAm) is a form of epigenetic regulation that is essential for human development and implicated in health and disease(Greenberg and Bourc'his 2019; Schübeler 2015). Through advancements in biological technology, large-scale DNA methylation profiling has become more affordable and widely used. Microarray technologies now enable the simultaneous interrogation of DNAm states of more than 850,000 CpG dinucleotides across the genome, using the latest EPIC array(Pidsley et al. 2016). An application of DNAm data has been in developing DNAm-based algorithms to predict health-related phenotypes, including blood cell type proportions(Houseman, Molitor, and Marsit 2014; Salas et al. 2018), ageing(Horvath 2013; Hannum et al. 2013; Q. Zhang et al. 2019; Horvath et al. 2018; Lin et al. 2016; Weidner et al. 2014; Lu, Seeboth, et al. 2019; Vidal-Bralo, Lopez-Golan, and Gonzalez 2016), all-cause mortality risk(Levine et al. 2018a; Lu, Quach, et al. 2019; Y. Zhang et al. 2017; Chen et al. 2016), cancer risk(Yang et al. 2016; Youn and Wang 2018), body-mass-index (BMI), and smoking signatures(McCartney et al. 2018), among others. These molecular predictors have great potential for clinical applications. A thorough and systematic investigation of their performance has however not been conducted so far.

Unlike the genome, the DNA methylome is of dynamic nature and largely explained by non-shared individual environments(Hannon et al. 2018). Like other high-throughput molecular data, DNAm can furthermore be impacted by variation in laboratory conditions, sample handling, reagents and/or equipment used(Leek et al. 2010). Technical variation is often widespread and tackling such effects is of critical importance to study biological variation in any -omic analysis, including DNAm. Over the years, a plethora of methods have been developed to identify and remove unwanted technical variations from DNAm data(Pidsley et al. 2013; Fortin et al. 2014; Xu et al. 2016; Teschendorff et al. 2013; Niu, Xu, and Taylor 2016; Xu et al. 2017; Maksimovic, Gordon, and Oshlack 2012). Previous studies have investigated the impact of specific methods on outcomes of DNAm analysis and demonstrated the importance of correcting for probe design type, batch effects, and hidden confounders while the effect of different normalization strategies gave mixed results(van Rooij et al. 2019; Wu et al. 2014; Wang et al. 2015; Marabita et al. 2013). A systematic and unbiased evaluation of commonly used data preprocessing and normalization strategies of DNAm data for the application of DNAm-based predictors has however not yet been conducted. DNAm is an important tool to study health and disease and understanding how analytical strategies impact algorithm performance is critical for method standardization and implementation for both research and clinical purposes.

Here, we performed a comprehensive investigation of 41 DNAm predictors and evaluated algorithm performance by measuring their test-retest reliability across 101 data preprocessing and normalization strategies in the Jackson Heart Study (JHS)(Taylor et al. 2005). The JHS has collected a large sample of 850K EPIC DNAm arrays in blood that includes 146 pairs of technical replicates. These replicates represent identical DNA samples that were assayed twice at independent time points. The agreement in DNAm predictor estimate between technical replicates after data preprocessing and normalization allowed us to quantify the degree to which an analytical

strategy can successfully remove unwanted technical variation. We report the best test-retest reliability for each predictor and demonstrate how reducing technical variation is critical for optimal algorithm performance in downstream phenotypic analyses. Our work emphasizes the importance of data processing and normalization of DNAm data and provides best practices to optimize the performance and reliability of DNAm predictors.

Methods

Cohort descriptions

The Jackson Heart Study is a large observational study of African American individuals from the Jackson, Mississippi (USA), metropolitan area(Taylor et al. 2005). JHS seeks to study the causes and disparities in cardiovascular health and related phenotypes in African Americans. Data and biological materials have been collected from 5,306 participants. For a subset of the cohort, peripheral blood samples were collected at baseline and subsequently used to quantify DNA methylation using the Illumina Infinium MethylationEPIC BeadChip that covers over 850,000 CpG sties. These samples have been included in previous DNAm studies(Lee et al. 2019; Lu, Quach, et al. 2019). See Table S1 for cohort characteristics. In our analysis, we included individuals for which DNAm data, phenotypic variables, and mortality data were available (N=1,909, 62.2% women, mean (SD) of age = 56.1 (12.4) years). For 146 individuals, technical replicates were collected. We therefore divided this dataset into two samples; 1) a general cohort sample that does not include technical replicate pairs (N=1,761, 62.6% women, mean(SD) of age=56.0 (12.3)) and 2) a technical replicate sample (N=146, 57.5% women, mean (SD) of age=57.4 (14.0)). Replicate pairs represent DNAm samples that were assayed twice using the EPIC array at separate occasions but originate from the same DNA extraction sample.

Data preprocessing and normalization strategies

To perform a systematic evaluation of available data preprocessing and normalization strategies, we incorporated all methods that are available through the commonly used R packages minfi(Aryee et al. 2014), wateRmelon(Pidsley et al. 2013), and ENmix(Xu et al. 2016). Within the same package, we implemented all possible combinations of background correction, dye-bias correction, probe correction, and data normalizations as was feasible within the structure of the package. In total, this yielded 101 strategies to prepare DNAm data (Table S2). For each sample, raw intensity values were read from IDAT files into an RGChannelSetExtended object in the R programming environment using the read.metharray() function in minfi. Sample quality control was performed by excluding samples with more than 5% of CpG sites with a detection P-value greater than 0.05 (using the pfilter() function in the wateRmelon package) and by removing outlying samples based on a low median of chipwide (un)methylation across CpG sites (using the getQC() function in minfi). In total, 44 samples were removed. No probes were filtered out to minimize missing probes in downstream DNAm prediction analysis. Data processing and normalization were then executed in batches of 96 samples for computational efficiency. The output of each analytical pipeline was a matrix with beta values for each sample. Table S3 shows an overview of our sample quality control analysis.

DNAm-based predictors

DNAm predictor estimates were calculated using regression coefficients as reported by the corresponding study unless stated otherwise. Custom R scripts were implemented that take as input a matrix of EPIC array beta values and output predicted estimates as a linear combination of weighted CpG methylation levels. For DNAm clocks, inverse transformation was applied to calibrate the DNAm age estimates in units of years, as required by the algorithm. For instance, Horvath's epigenetic clock regressed log-linear age (that leveraged age at 20) on DNA methylation levels and required this calibration step.

Next, we briefly describe the different predictors included in our study. Table S4 presents an overview of predictor characteristics. For full details on each predictor, we refer to their corresponding studies.

DNAm clocks:

The following predictors all output a form of DNAm age and capture a different aspect of biological age depending on characteristics of their training dataset. The Hannum clock uses 71 CpG probes and was developed in a whole blood 450K DNAm dataset of 656 individuals(Hannum et al. 2013). The Horvath clock was developed using 3,931 multi-tissue and -cell type samples using both 27K and 450K array samples(Horvath 2013). The Horvath clock uses 353 CpG probes that are present on both arrays. The BioAge4HAStatic clock is an extended measure of the Hannum clock and defined by forming a weighted average of Hannum's estimate with 3 cell types that are known to change with age: naïve (CD45RA+CCR7+) cytotoxic T cells, exhausted (CD28-CD45RA-) cytotoxic T cells, and plasmablasts(Chen et al. 2016). The Weidner clock uses 3 CpG and was developed in a 27K DNAm dataset of whole blood samples from 575 individuals(Weidner et al. 2014). The Lin clock uses 99 CpG and was developed in a dataset of 450K array whole blood samples of 656 individuals(Lin et al. 2016). The VidalBralo clock uses 8 CpG probes and was developed in a dataset of 450K array whole blood tissue of 390 individuals(Vidal-Bralo, Lopez-Golan, and Gonzalez 2016). The Skin & Blood clock uses 391 CpG probes and was developed in a dataset of 450K and EPIC arrays of a mixture of human fibroblasts, skin tissue, buccal cells, endothelial cells, whole blood, and cord blood samples (N=896)(Horvath et al. 2018). The Zhang clock uses 514 CpG probes and was developed in a dataset of EPIC and 450K arrays of 13,566 samples. The majority of the samples were derived from whole blood with a small subsample from saliva tissue(Q. Zhang et al. 2019).

Mitotic clocks:

The MiAge calculator uses 268 CpG probes and was developed on 4,020 samples of 8 cancer types using 450K DNAm arrays(Youn and Wang 2018). MiAge outputs an estimate of mitotic age (total number of lifetime cell divisions) for a given human tissue. The epiTOC calculator was developed in a 450K DNAm dataset of 650 whole blood samples. EpiTOC uses a subset of 385 Polycomb group targets promoter CpGs to predict an estimate of age acceleration in cancer. EpiTOC yields a score, denoted "pcgtAge", as the average DNAm over CpG sites, representing the age-cumulative increase in DNAm at these sites due to putative cell-replication errors(Yang et al. 2016).

Mortality risk estimators:

The Zhang mortality score is defined by a weighted average of 10 CpGs that are associated with mortality status(Y. Zhang et al. 2017). The Zhang mortality score predictor was trained on a discovery cohort of whole blood 450K DNAm samples from 954 individuals (N=402 deceased at follow-up) and validated in a cohort of 1,000 individuals (N=231 deceased at follow-up). The second mortality estimator, Levine clock, is a predictor of "phenotypic age", which is a DNAm surrogate of the composite score based on ten mortality markers (9 clinical markers + chronological age) (Levine et al. 2018b). A training cohort of 456 whole blood samples were then used to identify 513 CpGs predictive of phenotypic age. Only probes available on the 27K, 450K, and the EPIC array platform were used in their analysis. The linear combination of the weighted 513 CpGs is called "DNAm PhenoAge". The third mortality risk estimator isGrimAge from Lu et al., which is defined by a composite score based on seven DNAm-based plasma protein markers, DNAm-based pack years of smoking, chronological age and gender(Lu, Quach, et al. 2019). GrimAge used a training dataset of whole blood samples of 1,731 individuals. The DNA methylation profiling was based on the 450K beadchip but the biomarker was trained on the CpGs present on both the 450K and the EPIC array in order to ensure compatibility for both platforms. GrimAge was calculated using a python executable that was developed by the authors of the original study, which also outputs several DNAm-based plasma protein markers, three blood cell types, and pack years of smoking (see below).

Plasma protein markers:

DNAm-based estimators were developed for the following seven plasma proteins; adrenomedullin (ADM), beta-2-microglobulin (B2M), Cystatin-C, growth differentiation factor 15 (GDF-15), leptin, plasmin activator inhibitor 1 (PAI-1), tissue inhibitor metalloproteinases 1 (TIMP-1). These plasma proteins were measured using an immunoassay and the predictor trained using a whole blood 450k DNAm dataset of 1,731 individuals in Framingham Heart Study (FHS) cohort(Lu, Quach, et al. 2019). ADM, B2M, cystatin-C, GDF-15, leptin, PAI-1, and TIMP-1 are defined by 186, 91, 87, 137, 187, 211, and 42 CpGs, respectively. Each of these individual estimates were calculated using the GrimAge python executable.

Smoking predictors:

Two DNAm-based smoking predictors were included in our analysis. The Lu estimator was trained using a whole blood 450K DNAm dataset of 1,731 individuals in FHS and uses 172 CpGs for prediction, which is a component of GrimAge(Lu, Quach, et al. 2019). We estimated Lu pack years of smoking using the GrimAge python executable. The McCartney estimator was developed using EPIC DNAm data (only probes that are also present on the 450K platform) of 3,444 individuals(McCartney et al. 2018). The McCartney estimator uses 233 CpGs and outputs, similar to the Lu predictor, the number of pack years of smoking.

Blood cell type estimator:

We included DNAm-based blood cell type estimators for nine cell types in our analysis. For neutrophils (Neu), B cells, monocytes (Mono), natural killer cells (NK), CD4+ T cells (CD4T), and CD8+ T cells (CD8T), estimators were developed using 850K EPIC DNAm data from magnetic sorted cells(Salas et al. 2018). These six cell types were estimated using the estimateCellProp(refdata="FlowSorted.Blood.EPIC", nprobes=50) function of the ENmix R package. Plasma B cells (PlasmaBlasts), naive CD8+ T cells, and CD8+, CD28-,CD45RA- T cells (CD8pCD28nCD45RAn), were estimated based on the Horvath method (Horvath and Levine 2015) and computed using the same python executable as was used for the GrimAge estimator. These estimates are the same estimates that can be obtained through the online DNAm Age Calculator; https://dnamage.genetics.ucla.edu/.

Other estimators:

We also included DNAm-based estimators that are developed for body-mass-index (BMI, in kg/m2), alcohol (units: per week), educational attainment (Edu, in years), total cholesterol (in mmol/L), HDL cholesterol (in mmol/L), LDL with remnant cholesterol (in mmol/L), total:HDL cholesterol ratio (HDL_ratio), waist-to-hip ratio (WHR), body fat (in %). These estimators were developed in a whole blood EPIC DNAm dataset (only probes that are also present on the 450K platform) of between 2,819 to 5,036 individuals and used between 205 to 1,109 CpG sites to predict DNAm-based estimates(McCartney et al. 2018). Finally, we also included an estimator of leukocyte telomere length (TL). This DNAm-based TL predictor was developed in a whole blood 450K/EPIC DNAm dataset of 2,256 individuals and uses 140 CpGs(Lu, Seeboth, et al. 2019).

Statistical analyses

In the sample of technical replicates, the intraclass correlation (ICC) was calculated using the ICC() function of the R psych package (v2.1.3). More specifically, we use ICC(2,1), which is a type of ICC that calculates reliability from a single-measurement using a two-way random effects model(Shrout and Fleiss 1979; Koo and Li 2016). ICC(2,1) assumes absolute agreement, which means the estimates of the replicates are expected to have exactly the same value. We also calculated ICC(1,1), ICC(3,1), ICC(1,k), ICC(2,k), ICC(3,k) for comparison with other ICC types.

In the general JHS sample (i.e., without technical replicates), we calculated multiple statistical measures on the distribution of the output estimates of each predictor. The coefficient of variation was calculated by dividing the standard deviation by the mean of the distribution of the estimates. DNAm age acceleration residual (ΔAge) was calculated by regressing DNAm age on chronological age using the lm() function in R. To relate DNAm predictor estimates with mortality risk, a Cox proportional hazards regression model was fit using the coxph() function of the survival package (v3.2). Finally, to assess if the above statistical properties change depending on the type of data processing pipeline used, we calculated Spearman correlations between the ICC calculated in the replicate JHS sample and the various statistics generated in the general JHS sample across the 101 pipelines For this we use the cor.test(method="spearman") function of the stats package. The statistical analyses were performed in R (v4.0.3).

Results

To evaluate how unwanted technical variation in DNAm data impacts the performance of DNAm-based predictors, we implemented 101 data processing and normalization strategies in the JHS dataset. For each analytical strategy, which we will refer to as a "pipeline", we then extracted beta values and calculated estimates of 41 DNAm-based predictors in (1) JHS data1: a sample of 146 technical replicate pairs and (2) JHS data 2: a general sample of 1,761 non-replicate samples that do not overlap with the individuals in the replicate dataset. Figure 1 shows an overview of our analysis plan. In the sample of technical replicates, we quantified the average absolute agreement between replicate pair values (i.e. reliability) by means of the ICC for each DNAm predictor and each pipeline separately (41 predictors x 101 pipelines = 4,141 ICC analysis). We also generated DNAm estimates in the general sample. This allowed us to correlate the ICC of a pipeline that was estimated in the sample of replicates with predictor estimates in the independent general JHS sample.

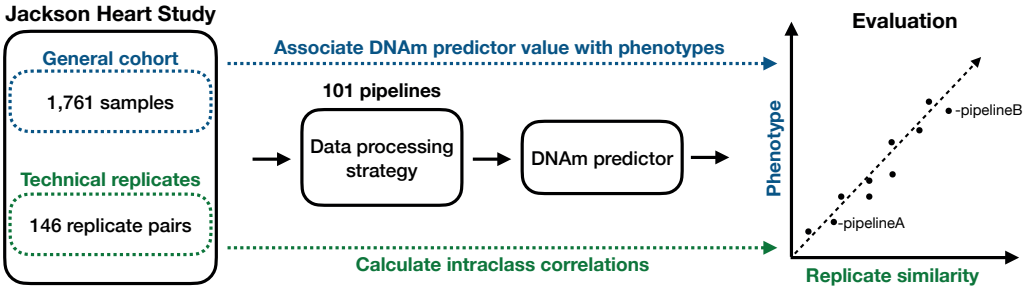


Figure 1. Schematic overview of analysis plan to evaluate DNAm algorithm performance. DNAm analyses are conducted using DNAm EPIC array samples in JHS. JHS includes a significant number of technical replicate pairs thereby allowing for a careful investigation of how the removal of unwanted technical variation impacts DNAm algorithm performance across 101 data processing pipelines. JHS has also collected information on disease-related phenotypes, including mortality status after follow-up. This allowed us to assess how removal of technical variation in DNAm predictor estimates by a data processing pipeline impacts downstream phenotypic association analyses.

We calculated the ICC estimates derived from a two-way random effect model to assess the reliability of each predictor for each data processing pipelines. The ICC is a zero to one estimate that quantifies the average absolute agreement across technical replicate pairs that were processed at a different occasion. We also calculated five other types of ICCs and found high concordance between the different ICC measures (mean rho=0.99, SD=0.01, see Figure S1). Table S5 reports all ICC statistics for each DNAm predictor and pipeline. In the remainder of the paper we will refer to ICC(2,1) as ICC, unless stated otherwise.

Predictor Information			Reliability (ICC statistics)			
Name	**Phenotype**	**Array compatibility**	**Median**	**Min**	**Max**	**Best pipeline**
GrimAge [15]	Mortality	EPIC/450K	0.990	0.921	0.994	ENmix: oob_mean_q3_rcp
ZhangAge [8]	Chronological age	EPIC/450K	0.991	0.987	0.992	ENmix: neg_mean_q2_rcp
TIMP_1 [15]	TIMP-1 serum protein	EPIC/450K	0.988	0.973	0.992	ENmix: oob_relic_q2_rcp
Bcell [5]	B-lymphocyte cell fraction	EPIC	0.980	0.881	0.988	Minfi: illumina_nobg_normcontrol
Neu [5]	Neutrophil cell fraction	EPIC	0.984	0.973	0.987	ENmix: oob_mean_q2_rcp
B2M [15]	B2M serum protein	EPIC/450K	0.973	0.759	0.985	ENmix: oob_relic_q1_rcp
SkinBloodAge [9]	Chronological age	EPIC/450K	0.979	0.908	0.982	ENmix: neg_relic_q1_rcp
Smoking_Lu [15]	Smoking pack-years	EPIC/450K	0.971	0.889	0.981	ENmix: oob_nodye_nonorm_rcp
Smoking_McCartney [20]	Smoking pack-years	EPIC	0.975	0.942	0.979	Minfi: noob_dyecorr
HannumAge [7]	Chronological age	450K	0.972	0.834	0.978	ENmix: est_relic_nonorm_rcp
CD8T [5]	CD8+ T-cell fraction	EPIC	0.969	0.881	0.978	ENmix: neg_mean_q1_rcp
NK [5]	Natural killer cell fraction	EPIC	0.952	0.883	0.977	ENmix: neg_relic_q3_rcp
BioAge4HAStatic [17]	Chronological age	450K	0.966	0.826	0.975	ENmix: oob_relic_nonorm_rcp
Cystatin_C [15]	Cystatin C serum protein	EPIC/450K	0.954	0.829	0.973	ENmix: oob_nodye_q2_rcp
PhenoAge [14]	Mortality	EPIC/450K/27K	0.954	0.926	0.97	ENmix: neg_relic_q1_rcp
Mono [5]	Monocyte cell fraction	EPIC	0.953	0.865	0.968	Minfi: illumina_bg_normcontrol
DNAmTL [12]	Telomere length	EPIC/450K	0.952	0.912	0.965	ENmix: oob_relic_q1_rcp
HorvathAge [6]	Chronological age	450K/27K	0.950	0.867	0.964	WateRmelon: naten
CD4T [5]	CD4+ T-cell fraction	EPIC	0.959	0.951	0.964	ENmix: neg_nodye_nonorm_rcp
epiTOC [18]	Mitotic divisions	450K	0.911	0.498	0.962	ENmix: oob_mean_q2_rcp
Leptin [15]	Leptin serum protein	EPIC/450K	0.896	0.447	0.953	ENmix: oob_relic_q3_rcp
VidalBraloAge [13]	Chronological age	27K	0.945	0.922	0.952	ENmix: neg_mean_nonorm_rcp
MiAge [19]	Mitotic divisions	450K	0.884	0.348	0.947	WateRmelon: nanes
LinAge [10]	Chronological age	450K	0.930	0.878	0.939	ENmix: est_relic_nonorm_norcp
ADM [15]	ADM serum protein	EPIC/450K	0.900	0.756	0.938	ENmix: neg_mean_q3_rcp
WHR [20]	Waist-to-hip ratio	EPIC	0.878	0.634	0.925	ENmix: oob_relic_q2_rcp

ZhangMortality [16]	Mortality	450K	0.877	0.807	0.92	Minfi: illumina_nobg_normcontrol
BodyFat [20]	Body fat	EPIC	0.893	0.843	0.918	ENmix: est_relic_nonorm_rcp
Cholesterol [20]	Total cholesterol	EPIC	0.888	0.762	0.917	ENmix: oob_nodye_q2_rcp
BMI [20]	BMI	EPIC	0.904	0.877	0.914	ENmix: neg_mean_nonorm_rcp
GDF_15 [20]	GDF-15 serum protein	EPIC/450K	0.819	0.502	0.903	ENmix: est_mean_q1_rcp
LDL [20]	LDL	EPIC	0.846	0.732	0.901	ENmix: oob_relic_nonorm_rcp
HDLratio [20]	Total:HDL cholesterol ratio	EPIC	0.848	0.643	0.890	ENmix: oob_relic_q1_rcp
Alcohol [20]	Alcohol	EPIC	0.807	0.551	0.878	ENmix: neg_relic_nonorm_rcp
WeidnerAge [11]	Chronological age	27K	0.826	0.583	0.865	ENmix: neg_relic_nonorm_rcp
Education [20]	Educational attainment	EPIC	0.774	0.506	0.865	Cross: noob_dyecorr_BMIQ
HDL [20]	HDL cholesterol	EPIC	0.835	0.694	0.853	ENmix: est_relic_q1_rcp
CD8pCD28nCD45Ran [6]	Specific T-cell fraction	27K	0.814	0.756	0.845	ENmix: oob_relic_nonorm_rcp
PlasmaBlast [6]	Plasma B cell fraction	27K	0.718	0.638	0.840	Cross: noob_dyecorr_BMIQ
PAI_1 [15]	PAI-1 serum protein	EPIC/450K	0.744	0.22	0.838	ENmix: neg_relic_q3_rcp
CD8naive [6]	CD8 T-cell fraction	27K	0.777	0.659	0.830	WateRmelon: danen

Table 1. Overview of predictor reliability and best performing data processing pipelines. *Shown are general information on each DNAm-based predictor alongside their corresponding ICC statistics from the reliability analysis. The name of the predictor, the phenotype it is trained on, and DNAm array compatibility are listed on the left side of the table. ICC statistics are listed on the right side of the table. For each predictor, across 101 pipelines, the median, minimum, and maximum ICC are the best performing data processing pipelines (i.e., the pipeline with listed. Predictors are ranked by the maximum ICC. The final column reports the highest reliability).*

Most DNAm-based predictors yield high reliability when the best analytical pipeline is implemented

Table 1 shows all 41 DNAm predictors alongside general information on each algorithm and corresponding ICC statistics, including the data processing and normalization pipeline that yielded the highest reliability for each predictor. Across all predictors and pipelines (N=4,141), we observed a significant degree of similarity between replicates (all ICC P-values < 0.05/4,141). The median across all ICC estimates is 0.93 with a range of 0.22-0.99.

The GrimAge predictor reports the highest reliability (ICC=0.994, P=6.6e-144), followed by ZhangAge (ICC=0.992, P=8.4e-132), and TIMP_1 (ICC=0.992, P=8.5e-133). In fact, 32 out of 41 predictors (78%) reach a reliability of an ICC > 0.9 with at least one data processing pipeline. The predictors with higher ICCs have more narrow ICC distributions than predictors with lower ICCs (see Figure 2), suggesting that predictors with higher reliability are more robust to the choise of data processing pipelines. The predictors with the lowest reliability are CD8pCD28nCD45RAn (ICC=0.85, P=1.63e-41), PlasmaBlast (ICC=0.84, P=7.19e-52), PAI-1 (ICC=0.84, P=2.80e-40), and CD8_naive (ICC=0.83, P=1.17e-39).

Across pipelines and predictors (N=4,141), the ENmix package yielded higher reliability (median ICC=0.93, range=0.61-0.99) than the minfi (median ICC=0.91, range=0.22-0.99) and wateRmelon (median ICC=0.91, range=0.49-0.99) packages. Among the best performance of each 41 DNAm predictors, i.e. achieving the highest reliability, 32 (78%), 4 (10%), and 3 (7%) predictors were from the ENmix, minfi, and wateRmelon package, respectively. Among ENmix pipelines; out-of-band (OOB) background estimation (15 out of 32), REgression on Logarithm of Internal Control probes (RELIC) dye-bias correction (19 out of 32), no quantile normalization (12 out of 32), and the Regression on Correlated Probes (RCP) probe-type bias correction (31 out of 32) yielded the highest reliability most often (see Figure S2). Two ENmix pipelines achieved the highest reliability for three predictors. The analytical pipeline that included OOB background estimation, RELIC dye-bias correction, no normalization, and RCP probe-type bias correction (i.e. "ENmix:oob_relic_nonorm_rcp") performed best for the BioAge4HAStatic, LDL, and CD8pCD28nCD45RAn predictors. The pipeline that included OOB background estimation, RELIC dye-bias correction, quantile normalization, and RCP probe-type bias correction (i.e. "ENmix: oob_relic_q1_rcp") performed best for the B2M, DNAmTL, and HDLratio predictors.

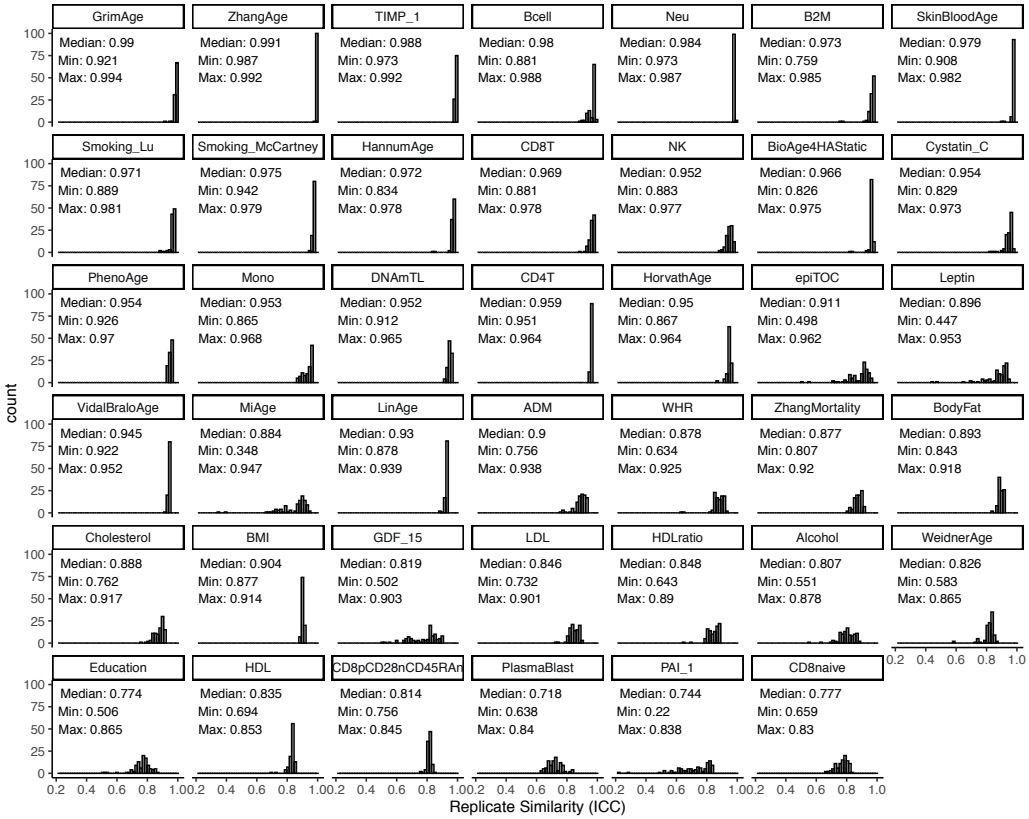


Figure 2. The distribution of intraclass correlations across pipelines for each DNAm algorithm. For each predictor, a histogram of ICC values across 101 pipelines is shown. The ICC quantifies the degree of absolute agreement between estimator values of a pair of technical replicates. The predictors are ranked based on their max ICC value. The name of the predictor is printed on top. In each panel, the median, lowest, and highest ICC value of a corresponding data processing pipeline for that predictor is shown as well.

There is significant heterogeneity in pipeline performance across predictors

Among the 41 best performing pipelines (i.e. the pipeline with the largest ICC value for each of the 41 predictors), there are 27 different data processing and normalization strategies, which highlights significant heterogeneity in choice of best pipeline between predictors. As ICC differences between pipelines of a predictor can be small and pipelines beyond the highest ICC may also be informative, we calculated the median rank across the 41 predictors for each of the 101 pipelines (see Table S6). The pipeline with the best median rank (at 15) across predictors is the "ENmix: oob_relic_q1_rcp". While this observation suggests this pipeline yields the best average performance across predictors, it still scored average to low for multiple predictors. For example, for the BMI predictor the "ENmix: oob_relic_q1_rcp" pipeline had one of the lowest ranks (ICC = 0.89, rank = 91). A data processing pipeline can also introduce more spurious variation instead of

removing technical variation. That is, the raw data pipeline that does not apply any data processing and normalization yielded a median rank of 85 (range: 7 to 100). For the CD4T and CD8 naive predictors, the raw data pipeline ranked as the seventh best performing pipeline highlighting that most pipelines perform worse than no data processing at all for these two predictors. The "Minfi: raw_quantile_strat" and "Minfi: illumina_bg_quantile_strat" had the lowest median rank of 100 and yielded the lowest reliability for 17 and 9 predictors, respectively (Table S5).

To assess the concordance in pipeline performance across predictors more formally, we calculated the rank correlation in pipeline reliability between all pairs of predictors. In Figure 3 we visualize the result of this analysis via a clustered correlation heatmap.

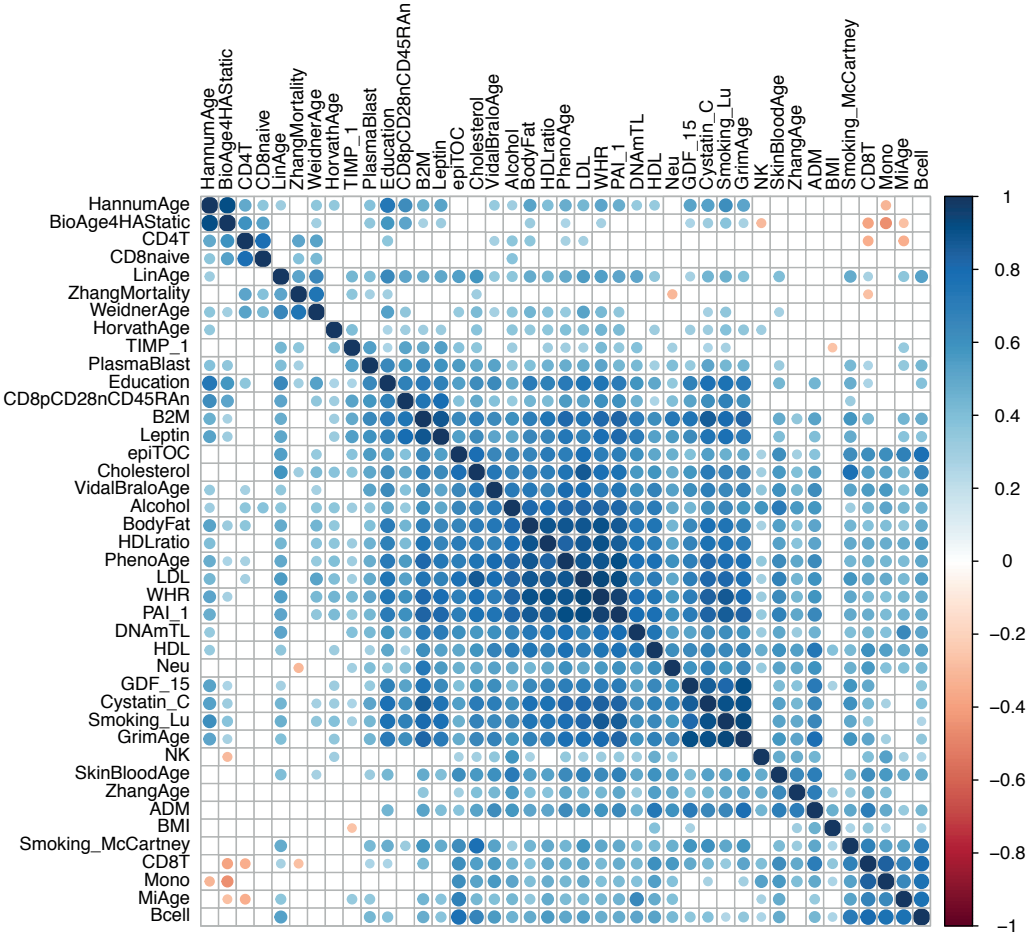


Figure 3. DNAm predictors have a moderate degree of concordance in performance between pipelines. Shown is a clustered correlation heatmap of pipeline reliability concordance between predictors. The color coding depicts Spearman's rho and clustering is performed using hierarchical clustering. Only correlations with a P-value < 0.01 are colored.

For some predictors the ranking in pipeline performance is very similar. For example, the GrimAge, Smoking_Lu, Cystatin_C, and GDF_15 predictors show strong concordance (mean rho = 0.92). As noted, these four predictors were developed in the same dataset and the Cystatin_C, GDF_15, and Smoking_Lu estimates are included in the GrimAge algorithm. Across all pairs of predictors, we find a moderate correlation in pipeline performance (mean rho=0.40, SD=0.27). Some predictors however show little to no concordance with other predictors. The ranking of pipelines of the BMI and NK predictor, for example, have a mean rank correlation of 0.14 (SD=0.20) and 0.21 (SD=0.24), respectively, with that of other predictors. For a handful of predictor-pairs we even observe a negative correlation, suggesting that pipelines that yield high reliability for one predictor yield low reliability for another. Pipeline performance of the BioAge4HAStatic and Mono predictors for example have a correlation of -0.45 (P=2.1e-06). Our findings thus far show that specific pipelines are more effective in removing unwanted technical variation for a predictor and that significant heterogeneity exists in pipeline performance across predictors.

The choice of data processing pipeline impacts downstream analysis of predictors

Next, we evaluated if the performance of a pipeline can also affect downstream phenotypic analyses of a predictor. For these analyses, we used the general JHS data 2 sample. For each pipeline, we calculated the mean and standard deviation (SD) of the predictor estimate distribution in the general JHS sample. For each predictor, we then correlated these two statistics (i.e., the mean and SD) with the ICC estimates of the pipelines obtained in the technical replicate sample. We find that the choice of pipeline has a significant impact on the distribution of the predictor estimate. Of the 41 predictors, 33 (80%) are significantly impacted on the distribution of their estimates after Bonferroni correction P<0.0012). For 22 predictors (54%), we find a significant correlation for both the mean and standard deviation. For DNAmTL, we, for example, observe a negative correlation between the performance of a pipeline and the mean of the estimate distribution (rho=-0.71, P<2.2e-16) and a positive correlation with the standard deviation of the estimate distribution (rho=0.79, P<2.2e-16) . The best performing pipeline yields a mean estimate of 6.83 kilobases (SD=0.34). The least performing pipeline yields a mean estimate of 7.20 kilobases (SD=0.29). This shows that the more effective a pipeline is in removing technical variation, the lower the DNAm-based predicted estimate of telomere length and the larger the variation between individuals. The direction of effect of the relationship between pipeline performance and the mean and standard deviation of the DNAm variables varies between predictors as well. HorvathAge, for example, is impacted on its standard deviation (rho=0.39, P=5.6e-05) but not on the mean (rho=-0.10, P=0.27). HDLratio is impacted on its mean but unlike DNAmTL shows a positive correlation with pipeline performance (rho=0.38, P=9.8e-05). HDLratio is not impacted on the standard deviation of its distribution (rho=0.00, P=0.96). Correlation plots and correlation statistics of all predictors are shown in Supplementary Note 1. A full overview of test statistics can be found in Table S7.

Several DNAm age predictors are known to predict all-cause mortality risk. We therefore examined if pipeline performance also impacts their association with mortality risk. We focus on four predictors: HorvathAge, PhenoAge, GrimAge, and ZhangAge. Each predictor has different training characteristics and captures a different aspect of biological age and/or mortality risk(Horvath and Raj 2018). ZhangAge is a blood-based DNAm clock and was developed

on the largest training dataset and shown not to be associated with mortality risk despite its improved precision(Q. Zhang et al. 2019). We find that pipeline performance significantly impacts downstream analysis for all four predictors (Figure 4).

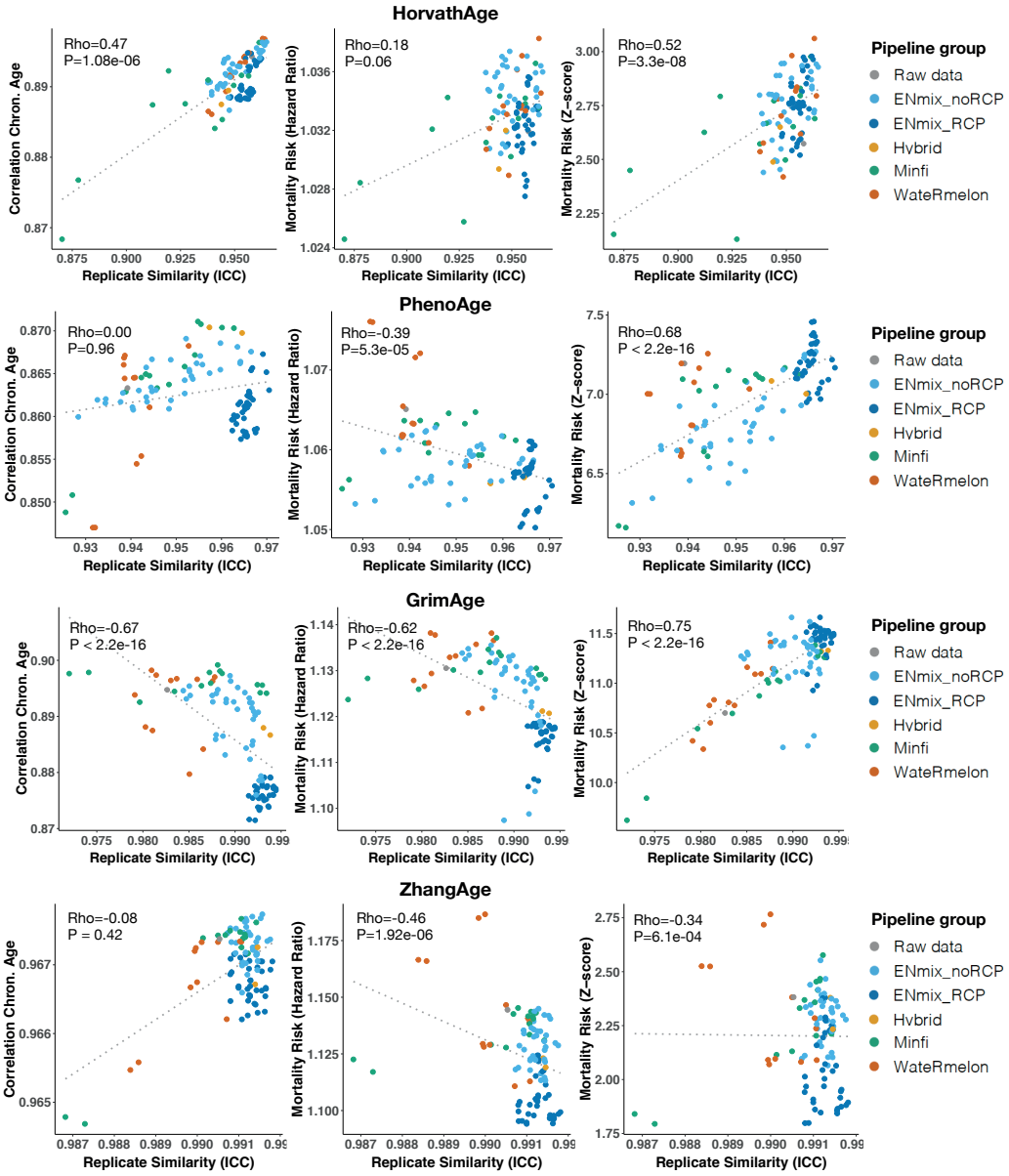


Figure 4. Pipeline performance impacts downstream analyses of DNAm age predictors. *Shown are association between pipeline ICC and the correlation with chronological age (left panels), the hazard ratio of mortality risk prediction (middle panel), and the z-score of the mortality risk prediction (right panels) for Horvath Age (top row), PhenoAge (2nd row), GrimAge (3rd row), and ZhangAge (bottom row). Pipelines are color-coded by package/ method. Spearman rank correlation statistics are shown in the top left corners.*

For HorvathAge, pipelines that achieve greater reliability also achieve a greater correlation between HorvathAge and chronological age (rho=0.47, P=1.8e06). Better performing pipelines furthermore achieve greater power to predict all-cause mortality (rho=0.52, P=3.3e-08). For PhenoAge, we did not find an effect on the correlation with chronological age but did find the survival analysis to be significantly impacted. Better performing pipelines achieve greater power for PhenoAge (rho=0.68, P<2.2e-16) but also a smaller hazard ratio (rho=-0.39, P=5.3e-05), suggesting that unsuccessful removal of technical variation in DNAm data can inflate the magnitude of mortality risk. In contrast to our findings for HorvathAge, we found that better performing pipelines produced a lower correlation with chronological age for GrimAge (rho=-0.67, P < 2.2e-16). Similar to PhenoAge, we found that pipelines that achieve greater reliability yield more significant associations with mortality for GrimAge (rho=0.75, P<2.2e-16) but also a smaller hazard ratio (rho=-0.62, P<2.2e-16). The most reliable pipeline reports a significant hazard ratio of 1.12 (SE=0.01, P=1.60e-30), which verifies GrimAge as a strong predictor of all-cause mortality, especially when spurious technical variation is appropriately accounted for. For ZhangAge, we found no impact on the correlation with chronological age. Better performing pipelines produced smaller and less significant effects in associations with all-cause mortality. The most reliable pipeline produced a non-significant hazard ratio of 1.10 (SE=0.05, P=0.06), confirming that ZhangAge does not predict mortality risk. Taken together, using the general JHS sample, we demonstrate how pipeline performance has a significant impact on downstream phenotypic analysis of DNAm predictors.

Predictor reliability is inversely associated with sample size of the training dataset

To assess if specific features of the predictors are associated with higher reliability, we investigated the number of CpG probes and the sample size of the training dataset in relation to the ICC of the best performing pipeline (see Figure S3). Using predictors for which such information was available, we find that the sample size of the dataset in which a predictor was developed is inversely associated with the observed predictor reliability (N=37, rho=-0.39, P=0.02). We did not find a significant association between the number of predictor CpG probes and reliability of a predictor (N=37, rho=-0.21, P=0.20).

A smaller number of replicate pairs can be used to measure reliability

In our analyses, we made use of a large number of replicate pairs. We therefore assessed how sample size affected our measure of reliability and if a smaller number of replicate pairs yield similar findings. Across reliabilities from all pipelines and predictors, we observe good concordance (rho > 0.94) with as low as ten replicate pairs compared with measures obtained from larger sample sizes (Figure S4). Differences however exist between predictors with some predictors still requiring a larger number of replicate pairs (Supplementary Note 2).

Discussion

DNAm-based predictors are emerging as powerful new methods to study health and disease, but little is known about the reliability of the estimates they produce. To investigate their performance, we carried out a systematic evaluation of 41 predictors across 101 data processing and normalization strategies and assessed to what degree algorithm performance is impacted by (un)successful removal of technical variation. Leveraging a large technical replicate sample in the JHS, we demonstrate that the choice of analytical pipeline has a significant impact on the reliability of predictors as well as on the outcomes of downstream phenotypic analyses. We highlight that specific pipelines are more effective in removing unwanted technical variation for a predictor but that significant heterogeneity exists in pipeline performance across predictors. Pipelines of the ENmix package achieved the highest reliability and were most frequently represented among the best performing pipelines. As research on DNAm-based predictors will continue to grow, our work provides best practices for the research community to help standardize their implementation and improve their performance.

To quantify method performance, we used a type of intraclass correlation that measures test-retest reliability by assessing the degree of absolute similarity between technical replicate pairs. Guidelines from reliability research suggest that ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability(Koo and Li 2016). The ICC range of best performing pipelines across predictors was 0.83-0.99, indicating good to excellent reliability for these predictors. For 32 out of 41 predictors (78%), we found excellent reliability (ICC > 0.9) for at least one data processing pipeline. Several predictors show a reliability close to 1, which demonstrates that repeated collections of DNAm data yield almost the same predictor estimate and highlights their potential as a biomarker for health-related outcomes. Among predictors with high reliability are predictors of mortality risk, smoking behavior, blood cell types, and cancer risk. Demonstrating internal validity for these DNAm tools is important for research purposes but even more so for their potential utilization for health management and disease prediction in the clinic. GrimAge, a strong predictor of all-cause mortality, for example, has the highest test-retest reliability of 0.994. This finding demonstrates excellent test-retest reliability based on technical replicates from the same biological sample. It remains an open question if the measured reliability translates to repeated measures of DNA samples extracted from different blood draws at the same time point or across time points. The analytical framework we applied can however be easily extended to study design of other types of (biological) replicates. Establishing method reliability in other contexts of technical and biological variation is an important next step for future research.

We found that the choice of analytical pipeline is essential as multiple data processing strategies produced poor reliability (ICC<0.5) for several predictors. For some predictors, like for CD4T and CD8 naive T cells, using the raw data achieves higher reliability than most data processing pipelines. This highlights that analytical decisions on how to best prepare DNAm data require careful consideration as certain data processing and normalization steps can even reduce algorithm performance. Among the best performing pipelines of each predictor, we found significant heterogeneity across predictors. That is, there are 27 unique pipelines across the

41 predictors. On average, pipelines of the Enmix package achieved the highest reliability most frequently. While there is no one optimal pipeline to use for all predictors, several data processing steps stand out as producing high reliability for multiple predictors. For example, almost half of the best performing pipelines make use of the RELIC dye-bias correction method. RELIC uses the information between pairs of internal normalization control probes to correct for differences between color channels that measure intensity levels of the array(Xu et al. 2017). The EPIC array contains 85 pairs of controls that target the same DNA region in housekeeping genes and contain no underlying CpG sites. RELIC uses the relationship between the pairs of controls to correct for dye-bias on intensity values for the whole array. Another data processing step that produced high reliability is the RCP probe type-bias correction method. 31 out of 41 of the best performing pipelines make use of this data processing step. RCP uses the existing correlation between pairs of nearby type I and II probes to adjust the beta values of all type II probes(Niu, Xu, and Taylor 2016). Both RELIC and RCP have been shown to reduce technical variation in DNAm data and are implemented in the ENmix package. While both approaches are effective in removing unwanted technical variation, we still recommend using the best performing pipeline for a specific predictor as reported in Table 1 as RELIC and RCP both show heterogeneity in performance across predictors.

The choice of analytical pipeline does not only impact the test-retest reliability of a predictor but also significantly affects downstream phenotypic analyses. We show that 80% of predictors are impacted on the mean and/or standard deviation of their distribution in the general JHS cohort. We furthermore analyzed DNAm clocks and showed that the strength of correlation between DNAm age and chronological age is affected in opposite directions for HorvathAge and GrimAge. While the correlation with chronological age becomes stronger with better performing pipelines for HorvathAge, the correlation becomes weaker for GrimAge. For DNAm clocks that are shown to be associated with mortality risk, successful removal of technical variation produced smaller hazard ratios but more significant associations. This highlights that not appropriately accounting for technical variation can decrease statistical power and inflate risk estimates for these predictors. It also shows that despite the narrow distribution of reliability estimates for these predictors, for example GrimAge has an ICC range of 0.921-0.994 indicating excellent reliability across all pipelines, the choice of pipeline still impacts downstream association analyses. We note that in our association analysis with mortality risk, we adjusted for chronological age, and still found that the choice of pipeline influences the outcome of the analysis. This is different from findings of a previous study that reported that the choice of pipeline influences the mean of DNAm age but not the DNAm age acceleration residual(McEwen et al. 2018). This study however only compared three data processing and normalization strategies and could have missed this effect as it did not perform a systematic evaluation across many pipelines. Finally, we confirm that ZhangAge, a DNAm clock developed in the largest blood based DNAm dataset, does not associate with mortality risk.

We also investigated if specific characteristics of a predictor impacted the measured reliability. We found that the sample size of the training dataset has a moderate inverse relationship with the reliability of a predictor. This suggests that predictors developed in larger training datasets are more sensitive to technical variation than predictors developed in a smaller dataset. This relationship could for example arise if larger training datasets on average have more technical

factors that are not properly accounted for. The ZhangAge predictor, however, was developed in the largest training dataset and shows the second to highest reliability of all predictors we investigated. This indicates that other factors in addition to sample size of the training dataset are likely to play a role as well. ZhangAge was developed using 65 training sets across 14 cohorts, where each training set had a certain number (ranging between 1 and 13) of cohorts randomly sampled from the 14 cohorts(Q. Zhang et al. 2019). This strategy is, as far as we know, unique to this predictor and may have helped select for CpG probes that are less impacted by technical variation due to its many training sets of different randomly assigned cohort compositions. As training datasets with large sample sizes are essential to developing more accurate DNA-based predictors, a strategy to randomize the potential effect of technical factors, like was implemented for the development of ZhangAge, could be worthwhile to consider for new predictors as well. We did not find a significant relationship between the number of CpG probes and the observed reliability of a predictor.

Our study comes with limitations. First, we measured reliability using technical replicate in one study. A different cohort or different types of repeated measures may yield different outcomes. Ideally, one would use study-specific replicate samples and assess if similar best practices are achieved or if alternative strategies are more appropriate to remove technical variation most optimally for that specific study. If future studies have the means to include replicate samples, they should aim to include at least ten replicate pairs. We determined that for most predictors a sample size of ten replicate pairs can already provide meaningful insights into their reliability. Second, several predictors were not fully compatible with the EPIC array platform. Predictors that were developed on older DNAm array platforms showed lower reliability. Missing probes could have affected the outcome of our analysis. Having said that, as the older 27K and 450K DNAm array platforms are discontinued, any future application of predictors that are not fully compatible with the EPIC array will face a similar challenge.

In summary, this study demonstrates that considerable variation exists in the performance of DNAm-based predictors depending on the data processing and normalization strategy implemented. Analytical pipelines that best remove unwanted technical variation in DNAm data achieve excellent test-retest reliability for most predictors thereby demonstrating their potential as biomarkers for health-related outcomes. DNAm is an important tool to study health and disease. As the number of DNAm predictors continues to rise, understanding how best to improve and implement these algorithms will be essential for downstream clinical applications.

Declarations

Ethics approval and consent to participate

All participants included in this study provided consent per procedures of the JacksonHeart Study.

Availability of data and materials

All DNA methylation and phenotypic information are available through the Jackson Heart Study. To submit a request for data, complete a data request form: https://www.jacksonheartstudy.org/Research/Study-Data/Data-Access

Competing interests

The authors declare that they have no competing interests.

Funding

ATL and SH were supported by NIH Grant 1U01AG060908 – 01.

Author contributions

APSO, SH, and RAO conceived of the study. APSO performed data analyses and primary writing of the manuscript. ATL developed the GrimAge executable and provided code to estimate the GrimAge predictors and run the survival analyses. SH and RAO oversaw the work. JGW and SH provided access to data of the JacksonHeart Study. All authors read, gave input on, and approved the final manuscript.

Acknowledgements

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

Disclaimers

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

References

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." Bioinformatics.

Chen, Brian H., Riccardo E. Marioni, Elena Colicino, Marjolein J. Peters, Cavin K. Ward-Caviness, Pei-Chien Tsai, Nicholas S. Roetker, et al. 2016. "DNA Methylation-Based Measures of Biological Age: Meta-Analysis Predicting Time to Death." Aging 8 (9): 1844–65.

Fortin, Jean-Philippe, Aurélie Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia Mt Greenwood, and Kasper D. Hansen. 2014. "Functional Normalization of 450k Methylation Array Data Improves Replication in Large Cancer Studies." Genome Biology 15 (12): 503.

Greenberg, Maxim V. C., and Deborah Bourc'his. 2019. "The Diverse Roles of DNA Methylation in Mammalian Development and Disease." Nature Reviews. Molecular Cell Biology 20 (10): 590–607.

Hannon, Eilis, Olivia Knox, Karen Sugden, Joe Burrage, Chloe C. Y. Wong, Daniel W. Belsky, David L. Corcoran, et al. 2018. "Characterizing Genetic and Environmental Influences on Variable DNA Methylation Using Monozygotic and Dizygotic Twins." PLoS Genetics 14 (8): e1007544.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." Molecular Cell 49 (2): 359–67.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." Genome Biology 14 (10): R115.

Horvath, Steve, and Andrew J. Levine. 2015. "HIV-1 Infection Accelerates Age According to the Epigenetic Clock." The Journal of Infectious Diseases 212 (10): 1563–73.

Horvath, Steve, Junko Oshima, George M. Martin, Ake T. Lu, Austin Quach, Howard Cohen, Sarah Felton, et al. 2018. "Epigenetic Clock for Skin and Blood Cells Applied to Hutchinson Gilford Progeria Syndrome and Studies." Aging 10 (7): 1758–75.

Horvath, Steve, and Kenneth Raj. 2018. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing." Nature Reviews. Genetics 19 (6): 371–84.

Houseman, Eugene Andres, John Molitor, and Carmen J. Marsit. 2014. "Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data." Bioinformatics 30 (10): 1431–39. Koo, Terry K., and Mae Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." Journal of Chiropractic Medicine 15 (2): 155–63.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael a. Irizarry. 2010. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." Nature Reviews. Genetics 11 (10): 733–39.

Lee, Yunsung, Dianjianyi Sun, Anil P. S. Ori, Ake T. Lu, Anne Seeboth, Sarah E. Harris, Ian J. Deary, et al. 2019. "Epigenome-Wide Association Study of Leukocyte Telomere Length." Aging 11 (16): 5876–94.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018a. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." Aging 10 (4): 573–91.

Lin, Qiong, Carola I. Weidner, Ivan G. Costa, Riccardo E. Marioni, Marcelo R. P. Ferreira, Ian J. Deary, and Wolfgang Wagner. 2016. "DNA Methylation Levels at Individual Age-Associated CpG Sites Can Be Indicative for Life Expectancy." Aging 8 (2): 394–401.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, et al. 2019. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan." Aging 11 (2): 303–27.

Lu, Ake T., Anne Seeboth, Pei-Chien Tsai, Dianjianyi Sun, Austin Quach, Alex P. Reiner, Charles Kooperberg, et al. 2019. "DNA Methylation-Based Estimator of Telomere Length." Aging 11 (16): 5895–5923.

Maksimovic, Jovana, Lavinia Gordon, and Alicia Oshlack. 2012. "SWAN: Subset-Quantile within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips." Genome Biology 13 (6): R44.

Marabita, Francesco, Malin Almgren, Maléne E. Lindholm, Sabrina Ruhrmann, Fredrik Fagerström-Billai, Maja Jagodic, Carl J. Sundberg, et al. 2013. "An Evaluation of Analysis Pipelines for DNA Methylation Profiling Using the Illumina HumanMethylation450 BeadChip Platform." Epigenetics: Official Journal of the DNA Methylation Society 8 (3): 333–46.

McCartney, Daniel L., Robert F. Hillary, Anna J. Stevenson, Stuart J. Ritchie, Rosie M. Walker, Qian Zhang, Stewart W. Morris, et al. 2018. "Epigenetic Prediction of Complex Traits and Death." Genome Biology 19 (1): 136.

McEwen, Lisa M., Meaghan J. Jones, David Tse Shen Lin, Rachel D. Edgar, Lucas T. Husquin, Julia L. MacIsaac, Katia E. Ramadori, et al. 2018. "Systematic Evaluation of DNA Methylation Age Estimation with Common Preprocessing Methods and the Infinium MethylationEPIC BeadChip Array." Clinical Epigenetics 10 (1): 123.

Niu, Liang, Zongli Xu, and Jack A. Taylor. 2016. "RCP: A Novel Probe Design Bias Correction Method for Illumina Methylation BeadChip." Bioinformatics 32 (17): 2659–63.

Pidsley, Ruth, Chloe C. Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C. Schalkwyk. 2013. "A Data-Driven Approach to Preprocessing Illumina 450K Methylation Array Data." BMC Genomics 14: 293.

Pidsley, Ruth, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. 2016. "Critical Evaluation of the Illumina MethylationEPIC BeadChip Microarray for Whole-Genome DNA Methylation Profiling." Genome Biology 17 (1): 208.

Rooij, Jeroen van, Pooja R. Mandaviya, Annique Claringbould, Janine F. Felix, Jenny van Dongen, Rick Jansen, Lude Franke, et al. 2019. "Evaluation of Commonly Used Analysis Strategies for Epigenome- and Transcriptome-Wide Association Studies through Replication of Large-Scale Population Studies." Genome Biology 20 (1): 235.

Salas, Lucas A., Devin C. Koestler, Rondi A. Butler, Helen M. Hansen, John K. Wiencke, Karl T. Kelsey, and Brock C. Christensen. 2018. "An Optimized Library for Reference-Based Deconvolution of Whole-Blood Biospecimens Assayed Using the Illumina HumanMethylationEPIC BeadArray." Genome Biology 19 (1): 64.

Schübeler, Dirk. 2015. "Function and Information Content of DNA Methylation." Nature.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." Psychological Bulletin 86 (2): 420–28.

Taylor, Herman A., Jr, James G. Wilson, Daniel W. Jones, Daniel F. Sarpong, Asoka Srinivasan, Robert J. Garrison, Cheryl Nelson, and Sharon B. Wyatt. 2005. "Toward Resolution of Cardiovascular Health Disparities in African Americans: Design and Methods of the Jackson Heart Study." Ethnicity & Disease 15 (4 Suppl 6): S6–4 – 17.

Teschendorff, Andrew E., Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. "A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 K DNA Methylation Data." Bioinformatics 29 (2): 189–96.

Vidal-Bralo, Laura, Yolanda Lopez-Golan, and Antonio Gonzalez. 2016. "Simplified Assay for Epigenetic Age Estimation in Whole Blood of Adults." Frontiers in Genetics 7 (July): 126.

Wang, Ting, Weihua Guan, Jerome Lin, Nadia Boutaoui, Glorisa Canino, Jianhua Luo, Juan Carlos Celedón, and Wei Chen. 2015. "A Systematic Study of Normalization Methods for Infinium 450K Methylation Data Using Whole-Genome Bisulfite Sequencing Data." Epigenetics.

Weidner, Carola Ingrid, Qiong Lin, Carmen Maike Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk Olaf Bauerschlag, et al. 2014. "Aging of Blood Can Be Tracked by DNA Methylation Changes at Just Three CpG Sites." Genome Biology 15 (2): R24.

Wu, Michael C., Bonnie R. Joubert, Pei-Fen Kuan, Siri E. Håberg, Wenche Nystad, Shyamal D. Peddada, and Stephanie J. London. 2014. "A Systematic Assessment of Normalization

Approaches for the Infinium 450K Methylation Platform." Epigenetics: Official Journal of the DNA Methylation Society 9 (2): 318–29.

Xu, Zongli, Sabine A. S. Langie, Patrick De Boever, Jack A. Taylor, and Liang Niu. 2017. "RELIC: A Novel Dye-Bias Correction Method for Illumina Methylation BeadChip." BMC Genomics 18 (1): 4.

Xu, Zongli, Liang Niu, Leping Li, and Jack A. Taylor. 2016. "ENmix: A Novel Background Correction Method for Illumina HumanMethylation450 BeadChip." Nucleic Acids Research.

Yang, Zhen, Andrew Wong, Diana Kuh, Dirk S. Paul, Vardhman K. Rakyan, R. David Leslie, Shijie C. Zheng, Martin Widschwendter, Stephan Beck, and Andrew E. Teschendorff. 2016. "Correlation of an Epigenetic Mitotic Clock with Cancer Risk." Genome Biology 17 (1): 205.

Youn, Ahrim, and Shuang Wang. 2018. "The MiAge Calculator: A DNA Methylation-Based Mitotic Age Calculator of Human Tissue Types." Epigenetics: Official Journal of the DNA Methylation Society 13 (2): 192–206.

Zhang, Qian, Costanza L. Vallerga, Rosie M. Walker, Tian Lin, Anjali K. Henders, Grant W. Montgomery, Ji He, et al. 2019. "Improved Precision of Epigenetic Clock Estimates across Tissues and Its Implication for Biological Ageing." Genome Medicine 11 (1): 54.

Zhang, Yan, Rory Wilson, Jonathan Heiss, Lutz P. Breitling, Kai-Uwe Saum, Ben Schöttker, Bernd Holleczek, Melanie Waldenberger, Annette Peters, and Hermann Brenner. 2017. "DNA Methylation Signatures in Peripheral Blood Strongly Predict All-Cause Mortality." Nature Communications 8 (March): 14617.

Supplemental Materials

Full supplemental materials can be found here:
https://www.biorxiv.org/content/10.1101/2021.09.29.462387v1.supplementary-material

	Replicate Sample	General Sample
Individuals (N)	146	1,761
Age (years)	57.4 (SD = 12.3)	56.1
Female (%)	62.6 %	62.2 %
Death (%)	-	15.1%
EPIC arrays (N)	292	1,761

Table S1. Cohort characteristic of JHS. In total, 1,907 individual and 2,053 850 EPIC array samples were included in our analysis. Our analyses were conducted in two subsets of this cohort, a subset that consist of only the technical replicate samples (n=146 individuals) and a subset that contains the remainder of the cohort. Shown above are standard cohort characteristics, including the percentage of individuals that died after follow-up for the general sample.

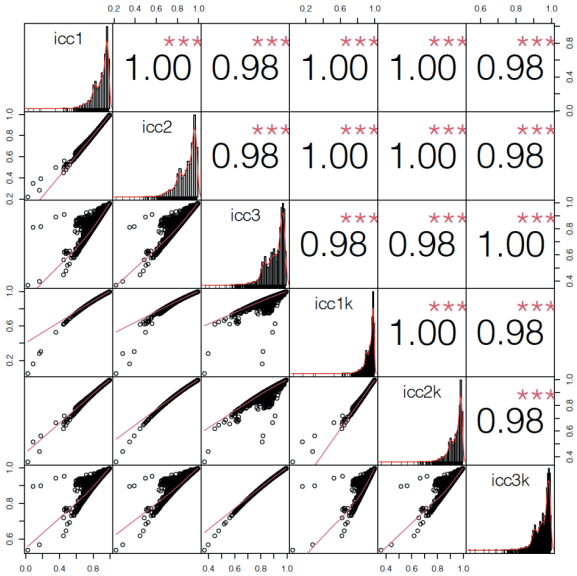


*Figure S1. Comparative analysis of ICC types across DNAm-based predictors and data processing pipelines. Shown are the bivariate scatterplots (left bottom) and the Spearman correlation (right top) between ICC types across all pipelines and predictors (N= 101x41 = 4141). The distribution of each ICC type is shown on the diagonal. ***P-values < 2.2e-16. This figure was made using the chart.Correlation() function of the PerformanceAnalytics R package (v2.0.4).*

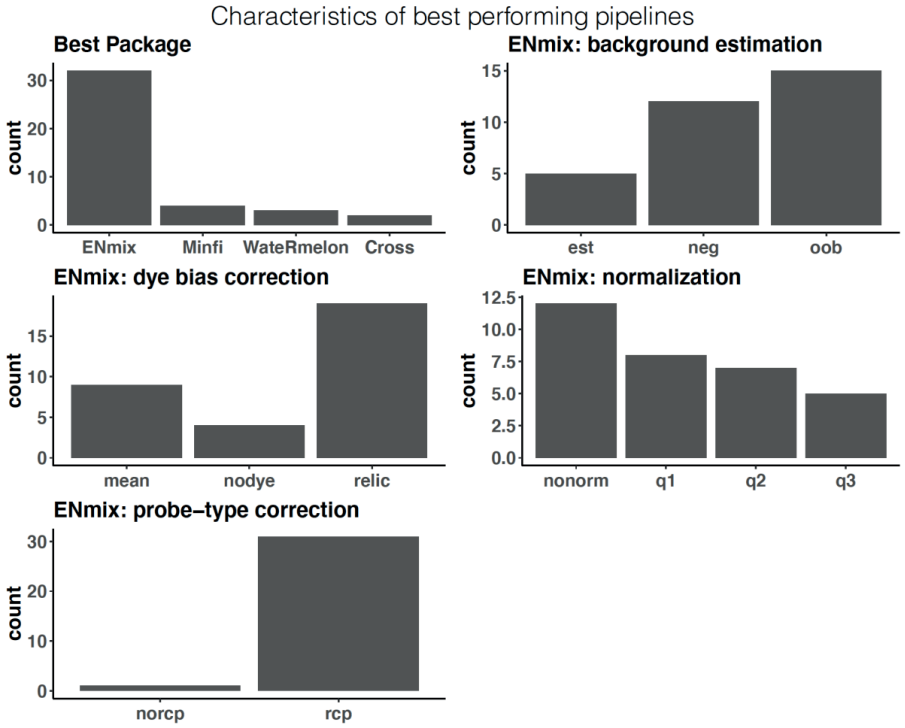


Figure S2. Characteristics of best performing pipelines of predictors. *These graphs are based on the 42 best performing data processing pipelines (i.e., pipeline with the highest reliability of each predictors). Top left shows the corresponding package. 32 out of 41 pipelines are part of the Enmix package. The top right shows which background estimations ranked among the 32 Enmix pipelines. Middle left shows the ENmix dye bias correction method. Middle right shows the Enmix normalization method. The bottom graph shows if a pipeline used probe-type bias correction (i.e. "RCP method").*

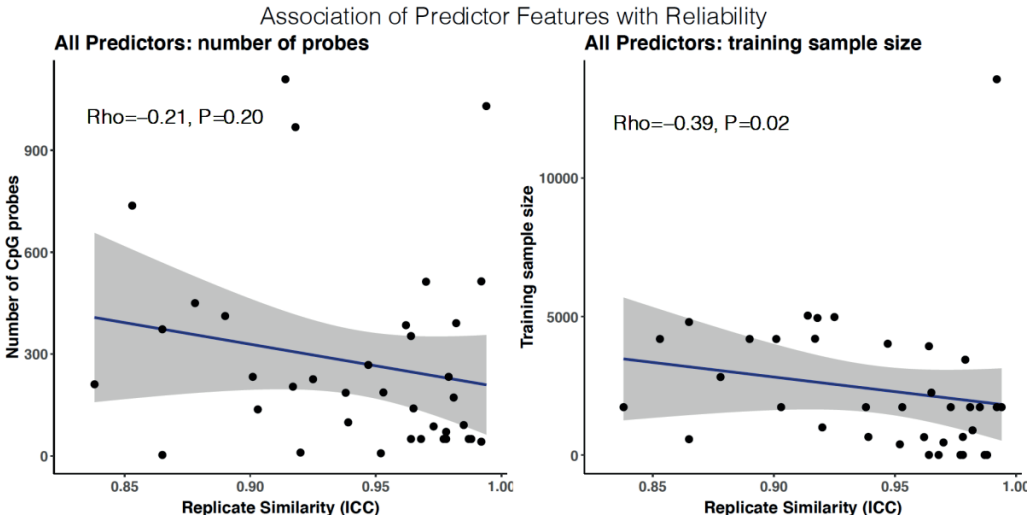


Figure S3. Association of predictor features with reliability. *Shown are scatter plots of the relationships between predictor features (i.e., training sample size and number of CpG probes) and the reliability (i.e., ICC) of the best performing pipeline for each predictor. Shown are the statistics of the correlation test (method="spearman") and a corresponding regression line.*

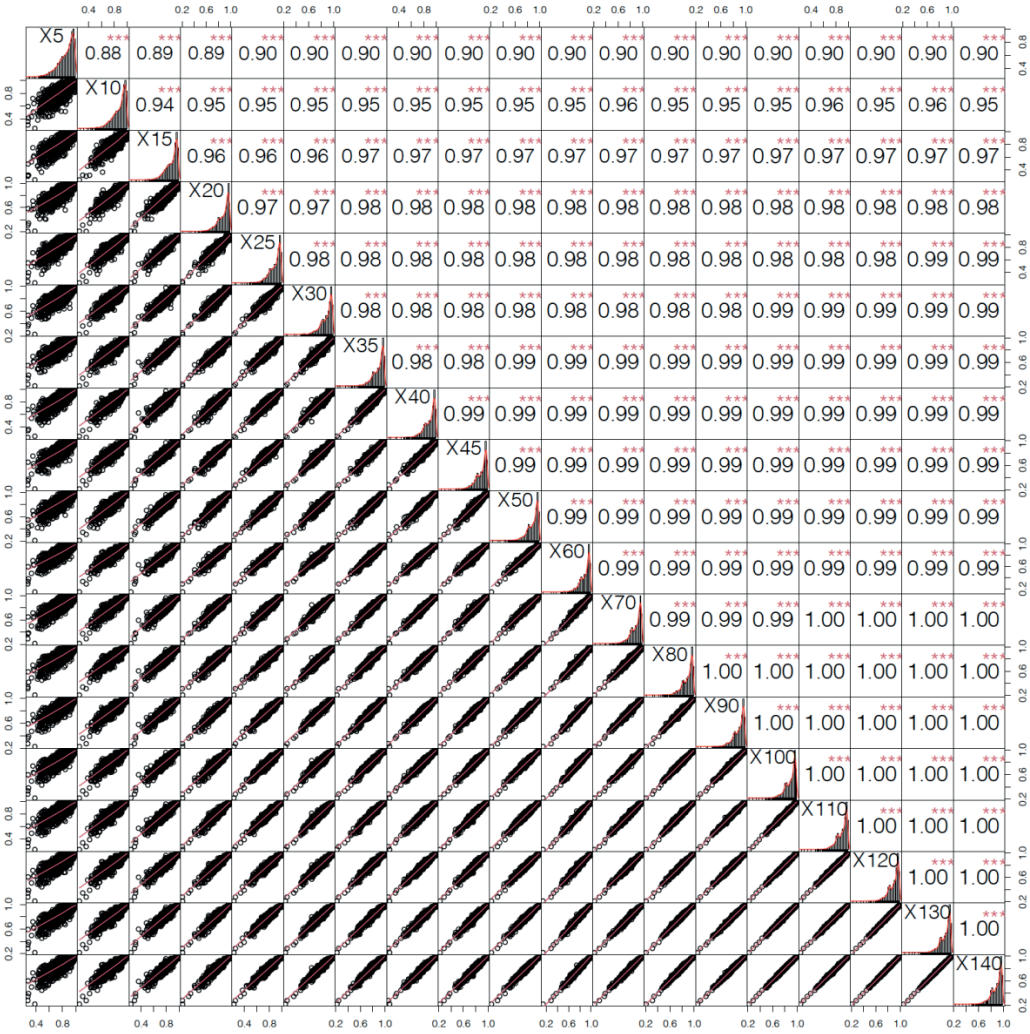


Figure S4. Reliability measures across different sample sizes of replicate pairs. *Shown are the bivariate scatterplots (left bottom) and the Spearman correlation (right top) between the interclass correlations (all pipelines and predictors (N= 101x41 = 4141)) obtained across different sample sizes of replicate pairs. The sample size of the set of replicate pairs is shown on the diagonal across. For each sample size, we performed a bootstrap analysis in which we randomly selected the specified number of pairs from the total of 146 replicate pairs and computed the intraclass correlation across ten independent samplings. We then computed the mean intraclass correlation across these ten samplings and correlated this obtained mean ICC across different sets of replicate pairs. ***P-values < 2.2e-16. This figure was made using the chart.Correlation() function of the PerformanceAnalytics R package (v2.0.4).*

CHAPTER 6

Epigenetic age is accelerated in schizophrenia with age- and sex-specific effects and associated with polygenic disease risk

Authors

Anil P.S. Ori[1]
Loes M. Olde Loohuis[1]
Jerry Guintivano[2]
Eilis Hannon[3]
Emma Dempster[3]
David St. Clair[4]
Nick J Bass[5]
Andrew McQuillin5
Jonathan Mill[3]
Patrick F Sullivan[2,7]
Rene S. Kahn[8]
Steve Horvath[9,10]
Roel A. Ophoff[1,10,11]

Affiliations

[1.] University of California Los Angeles, Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA
[2.] University of North Carolina, Department of Genetics, Chapel Hill, NC, US
[3.] University of Exeter, University of Exeter Medical School, Exeter, UK
[4.] University of Aberdeen, Institute of Medical Sciences, Aberdeen, Scotland, UK
[5.] University College London, Division of Psychiatry, UK
[6.] King's College London, London, UK
[7.] Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Stockholm, Sweden
[8.] Icahn School of Medicine at Mount Sinai, Department of Psychiatry, New York, NY, USA
[9.] University of California Los Angeles, Department of Biostatistics, Fielding School of Public Health, Los Angeles, CA, USA
[10.] University of California Los Angeles, Department of Human Genetics, David Geffen School of Medicine, Los Angeles, CA, USA
[11.] Erasmus University Medical Center, Department of Psychiatry, Rotterdam, The Netherlands

Abstract

The study of biological age acceleration may help identify at-risk individuals and contribute to reduce the rising global burden of age-related diseases. Using DNA methylation (DNAm) clocks, we investigated biological aging in schizophrenia (SCZ), a severe mental illness that is associated with an increased prevalence of age-related disabilities and morbidities. In a multi-cohort whole blood sample consisting of 1,090 SCZ cases and 1,206 controls, we investigated differential aging using three DNAm clocks (i.e. Hannum, Horvath, Levine). These clocks are highly predictive of chronological age and are known to capture different processes of biological aging. We found that blood-based DNAm aging is significantly altered in SCZ with age- and sex-specific effects that differ between clocks and map to distinct chronological age windows. Most notably, differential phenotypic age (Levine clock) was most pronounced in female SCZ patients in later adulthood compared to matched controls. Female patients with high SCZ polygenic risk scores (PRS) present the highest age acceleration in this age group with +4.30 years (CI: 2.40-6.20, P=1.3E-05). Phenotypic age and SCZ PRS contribute additively to the illness and together explain up to 22.4% of the variance in disease status in this study. This suggests that combining genetic and epigenetic predictors may improve predictions of disease outcomes. Since increased phenotypic age is associated with increased risk of all-cause mortality, our findings indicate that specific and identifiable patient groups are at increased mortality risk as measured by the Levine clock. These results provide new biological insights into the aging landscape of SCZ with age- and sex-specific effects and warrant further investigations into the potential of DNAm clocks as clinical biomarkers that may help with disease management in schizophrenia.

Manuscript status: submitted
Preprint available: https://doi.org/10.1101/727859

Introduction

As the population continues to age, reducing the burden of age-related disability and morbidity is timely and important, particularly for mental illnesses (Taylor and Reynolds 2020; Moffitt and Caspi 2019). Ranked as one of the most disabling illnesses globally(Salomon et al. 2015), schizophrenia (SCZ) has significant impact on patients, families, and society. SCZ is associated with a two- to threefold increased risk of mortality(McGrath et al. 2008; Olfson et al. 2015; Allebeck 1989a) and a 15 year reduction in life expectancy compared to the general population(Hjorthøj et al. 2017; Laursen, Nordentoft, and Mortensen 2014). Despite elevated rates of suicide and other unnatural causes of death, most morbidity in SCZ is attributed to age-related diseases, such as cardiovascular and respiratory diseases and diabetes mellitus(Saha, Chant, and McGrath 2007; Hayes et al. 2017; Olfson et al. 2015). Processes of biological aging may therefore be accelerated in patients diagnosed with SCZ, either through an increased prevalence of age-related conditions or as a more integrated part of the illness (Kirkpatrick et al. 2008). Quantification of biological aging can help with identification of at-risk individuals or even prevention of age-related diseases (Belsky et al. 2015; Field et al. 2018). While different aging biomarkers have been studied in SCZ, no clear demonstration of altered biological age has been shown (Nguyen, Eyler, and Jeste 2018). The recent development of DNA methylation (DNAm) age predictors however offers new opportunities to study the phenomenon of aging in SCZ.

DNAm age predictors, or "epigenetic clocks", are biomarkers of ageing that generate a highly accurate estimate of chronological age, known as DNAm age (Horvath 2013; Hannum et al. 2013; Levine et al. 2018). The difference (Δage) between predicted DNAm and chronological age is associated with a wide-range of health and disease outcomes, including all-cause mortality (Marioni et al. 2015; Chen et al. 2016; Perna et al. 2016; Levine et al. 2015), socioeconomic adversity and smoking(Fiorito et al. 2017), metabolic outcomes, such as body mass index (BMI) and obesity (Quach et al. 2017; Horvath et al. 2014), and brain-related phenotypes, such as Parkinson's disease, posttraumatic stress disorder, insomnia, major depressive disorder, and bipolar disorder (Horvath and Ritz 2015; Boks et al. 2015; Carroll et al. 2017; Han et al. 2018; Fries et al. 2017). As epigenetic signatures can be modifiable (Sugden et al. 2019), DNAm-based predictors may have significant clinical utility. Studies of DNAm aging so far found limited to no evidence for altered biological age in either brain or blood in SCZ (Voisey et al. 2017; Okazaki et al. 2019; Viana et al. 2017; McKinney et al. 2017). These studies, however, (i) consisted of small sample sizes and thus limiting the ability to detect a biological signal, (ii) used a single DNAm clock that may have not been most informative for aging studies of mental illnesses, and (iii) did not consider aging differences across the lifespan of patients. As morbidities in the SCZ population differ between older and younger individuals, and females and males (Olfson et al. 2015), analyses of both age- and sex-specific effects is warranted and could identify differential aging patterns, nevertheless.

To investigate DNAm aging in SCZ, we used three independent DNAm age estimators; the Hannum (Hannum et al. 2013), Horvath (Horvath 2013), and Levine clock (Levine et al. 2018). Each clock is designed using different training features and captures distinct characteristics of aging (Horvath and Raj 2018); (i) the Hannum age predictor was trained on whole blood adult samples, (ii) the Horvath predictor was trained across 30 tissues and cell types across developmental stages, and (iii) the Levine combines DNAm from adult blood samples with clinical blood-based

measures. As the Levine estimator is trained on chronological age and nine clinical markers, its output is referred to as DNAm PhenoAge or "phenotypic age". The Hannum estimator is said to capture measures of cell extrinsic aging in blood, whereas the Horvath clock measures more cell intrinsic aging as it was trained across multiple tissues and therefore is less dependent on cell type composition. All three clocks, in different but complementary ways, capture the pace of biological aging that is associated with various age-related conditions and diseases, including all-cause mortality (Horvath and Raj 2018; Chen et al. 2016).

DNAm clocks were implemented across four European case-control cohorts, representing a sample of almost twice the size of the largest SCZ DNAm age study conducted so far. Analyses are performed across the full sample and stratified by age and sex. We then integrated DNAm age with age of onset, duration of illness, and SCZ polygenic risk. DNAm smoking scores and blood cell type proportions were used to gain further insights into differential aging patterns. This study overall reports an in-depth investigation of the DNAm aging landscape in schizophrenia.

Material and Methods

Cohort and sample description

Details of samples included in this study can be found in the Supplementary Information. Briefly, unrelated patients with SCZ and ancestry-matched non-psychiatric controls from four cohorts of European ancestry were included; the Netherlands (N=1,116), Scotland (N=847), Sweden (N=96), and the United Kingdom (N=675). Cases were selected on the basis of a clinical diagnosis of SCZ using the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV), Research Diagnostic Criteria (RDC), or the International Classification of Diseases 10 (ICD10). Controls were unaffected subjects without a history of any major psychiatric disorder. Whole blood DNAm data was available for a total of 2,707 samples (1,399 cases and 1,308 controls; Table S1).

Genome-wide DNA methylation profiling and data processing

To quantify DNA methylation, DNA was extracted from whole blood and bisulfite converted for hybridization to the Illumina Infinium Human Methylation Beadchip. Samples were assayed with either the 27K or 450K beadchip, which contain 27,578 and 485,512 probes that interrogate CpG sites across the genome, respectively. For each platform, data processing pipelines were implemented, which includes background correction, color channel and probe type correction, and normalization of the data, to minimize the effect of technical variation on the final beta values. Samples with more than 5% of probes detected at $P > 0.05$ were excluded from further analyses (n=13). Full details are described in the supplementary methods.

DNAm-based estimation of biological age

To compute blood-based DNAm age estimates, processed beta values were used as input to the Hannum(Hannum et al. 2013), Horvath(Horvath 2013), and Levine (Levine et al. 2018) DNAm clock. These DNAm age estimators use a set of CpGs that are selected via an optimization algorithm to collectively minimize the error associated with estimating chronological age (Supplementary Information). Horvath DNAm age estimates were calculated using R scripts

from the Horvath DNA Methylation Calculator (https://dnamage.genetics.ucla.edu). Hannum and Levine estimates were obtained by using the reported set of probes with corresponding regression weights. We define Δage by subtracting chronological age at the time of the blood draw from the predicted DNAm age.

Statistical analyses

To investigate epigenetic aging differences in SCZ, we first removed samples with discrepant phenotypic sex and predicted sex based on DNAm data (n=9), as well as samples with missing chronological age data (n=237), bipolar disorder diagnosis (n=26), and duplicate samples (n=126). For each epigenetic clock, we regressed Δage on technical principal components (PCs), using the first components that cumulatively explain >90% of variation in intensity values of control probes, and added the residuals to mean(Δage) to generate a measure in the same units as Δage that is adjusted for technical variation (Δage-adjusted). We used the adjusted value for subsequent analyses and refer to it as Δage.

As association analyses of DNAm age between groups are sensitive to the distribution of chronological age, particularly at older ages, any case older than the oldest control was excluded from each cohort (n = 5 for NLD, 16 for SCT, 4 for SWD, and 1 for UK). Chronological age was furthermore included as a covariate in all analyses, as recommended(Khoury et al., n.d.). To minimize the effect of outlying samples, we excluded samples >3SD from mean Δage across cohorts (ranging from n=13 to 16 for the three clocks). These are samples for which DNAm age diverged substantially from chronological age, which are likely artifacts.

For each clock and each cohort, we implemented a multivariable regression model predicting Δage as a function of schizophrenia status, sex, and age. For the Dutch cohort, batch and array platform were also included as covariates, as this cohort consists of multiple datasets from both the 27K and 450K platform. For each clock, regression coefficients with corresponding standard errors for each of the four cohorts were then supplied to the rma() function of the metafor package(Viechtbauer 2010) in R to fit a meta-analytic fixed-effect model with inverse-variance weights and obtain an overall effect size and test statistic. To quantify the significance of age- and sex-specific effects, we determined the contribution of interaction effects on top of the main disease effect. We first combined all cohorts to maintain necessary sample sizes across age and sex groups. Age groups were defined by grouping samples by decades with ages 18 and 19 included in the first decade (18-30, 31-40, etc.). To quantify the gain in variance explained in Δage, models with the interaction term were compared to a baseline model without the interaction term. For each analysis, statistical significance was determined using Bonferroni correction, i.e. $P < 0.05$ / number of tests.

SCZ polygenic risk quantification

Polygenic risk scores (PRS) were obtained from analyses of the SCZ GWAS conducted by Psychiatric Genomics Consortium (PGC)(Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). Using a leave one out approach, weights were generated in a training dataset based on all samples minus the target cohort in which the PRS were calculated. For each individual, weighted single nucleotide polymorphisms (SNPs) were summed to a

genetic risk score that represents a quantitative and normally distributed measure of SNP-based SCZ genetic risk. To reduce between cohort-variation and maximize statistical power, we used a previously developed analytical strategy that uses principal component analysis (PCA) to concentrate disease risk across PRSs of ten GWAS p-value thresholds into the first principal component (PRS1)(Bergen et al. 2019) (Supplementary Information). PRS1 explains 70.7% of the variance in risk scores and 19.9% of the variance in SCZ status, which is more than any of the original p-value thresholds (4.9-17.4%). The other PCs had no explanatory value in disease status (mean $R^2 = 0.0\%$), which means that PRS1 captures the majority of SNP-based SCZ polygenic risk. PRS1 was generated for 1,933 individuals, 853 cases and 1080 controls, and modelled as both a quantitative and categorical variable to predict Δage.

Defining age at onset and illness duration

Age at onset is defined as the earliest reported age of psychotic symptoms or by the Operational Criteria Checklist (OPCRIT), depending on the cohort. This data is available for a subset of cases (N = 710) across the Dutch, Scottish, and UK cohorts. Illness duration is defined as the time between age at onset and blood collection. A more detailed description of each cohort's definition is available in the Supplementary Information.

DNA methylation-based smoking scores and blood cell type proportions

Smoking scores and blood cell type proportions were estimated from the data (see Supplementary Methods) and used as a proxy to further decompose differential aging effects.

Estimating the contribution of differential aging in schizophrenia

Using a multivariable logistic regression model for disease status, we fitted batch, cohort, DNAm smoking score, DNAm blood cell type proportions, and Δage as explanatory variables. We first performed a variable reduction step to select the most contributing variables to disease status by use of a regularized logistic regression using the glmnet() function in R ("glmnet" package, v2.13)(Friedman, Hastie, and Tibshirani 2010). Alpha was set to "1" (Lasso) and the lambda parameter estimated at the optimal value that minimizes the cross-validation prediction error rate using the cv.glmnet() function. For each selected variable, we then report the variance explained in SCZ status (glm, family = "binomial") for both the individual variable as well as adjusted for all other selected variables using the NagelkerkeR2() function in the "fmsb" package (v 0.6.3). The significance of each variable to their contribution was computed by comparing the model with and without the variable of interest using the likelihood ratio test of the anova() function.

Results

Figure 1 shows a schematic overview of the study design and analysis framework used to investigate DNAm aging in SCZ. After data pre-processing and quality control, 1,090 SCZ cases and 1,206 controls (2,296 subjects of 2,707 initial samples) were included in our analysis. The overall sample has a mean age of 40.3 years (SD=14.4) and consists of 34.5% women (Table S1 and Figure S1).

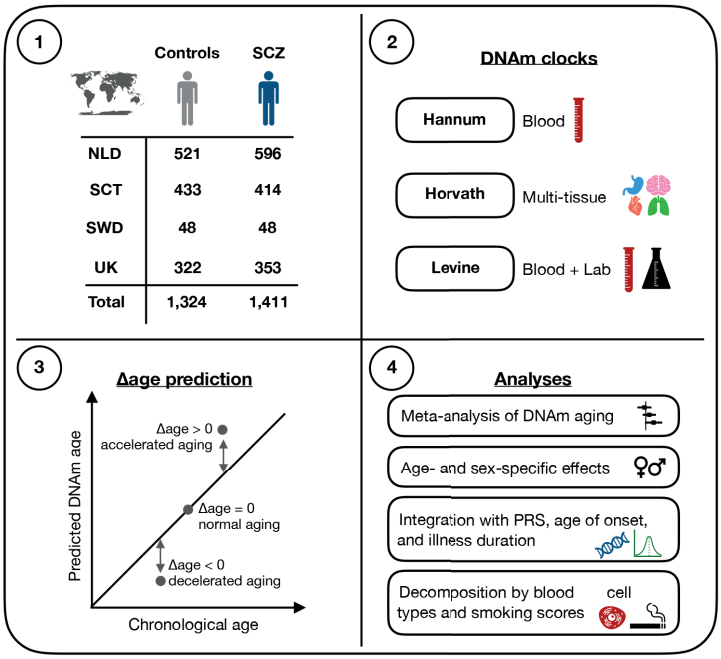


Figure 1. Overview of study design and analysis framework. *DNA methylation (DNAm) data was available for a total of 2,735 samples across four European cohorts. See Table S2 for more details on samples. DNAm age estimates were generated using three DNAm clocks, each designed to capture different features of aging (box 2). To investigate differences in aging between cases and controls, Δage was computed (box 3) and analyzed according to the step-wise framework shown in box 4. SCZ = schizophrenia, NLD=Netherlands, SCT=Scotland, SWD=Sweden, UK=United Kingdom, PRS=polygenic risk scores.*

Across cohorts, all three clocks produce a high correlation with chronological age (Pearson's r = 0.92-0.94; Figure 2A and S2). Using duplicates in the Dutch cohort, we assessed consistency between pairs of technical replicates, i.e. samples for which blood was collected at the same time but DNA processed at different times and DNAm data obtained on different arrays. Comparing Δage estimates between these pairs, we find a significant correlation for each clock (Figure S3); Hannum (rho = 0.79, n = 10), Horvath (rho = 0.53, n=118), Levine (rho = 0.67, n=118). Δage directionality (i.e. age deceleration or acceleration) is concordant in 90%, 73%, and 86% of pairs for Hannum, Horvath, and Levine, respectively, highlighting that the obtained estimates of DNAm age are reproducible for all three clocks. Comparing Δage estimates between clocks using all samples, we find a moderate concordance (Pearson's r = 0.39-0.43; Figure S4), demonstrating that a significant proportion of the variation in Δage is clock-specific. As these three estimators were trained on different features of biological aging, investigating them in conjunction may thus yield broader insights into differential aging.

DNA methylation age is altered in an age-dependent manner

Across the full sample, patients with SCZ are on average 1.53 years older in phenotypic Δage (Levine clock) compared to controls (Pmeta= 3.45E-08) (Figure 2B). The intrinsic cellular age (Horvath) predictor revealed an opposite pattern, with SCZ cases appearing 0.47 years younger compared to controls (Pmeta= 0.06). No differences were observed between cases and controls when applying the blood-based Hannum DNAm age predictor. Within the analysis of each clock, we observed no evidence of heterogeneity between the four cohorts (Phet > 0.05, Table S5).

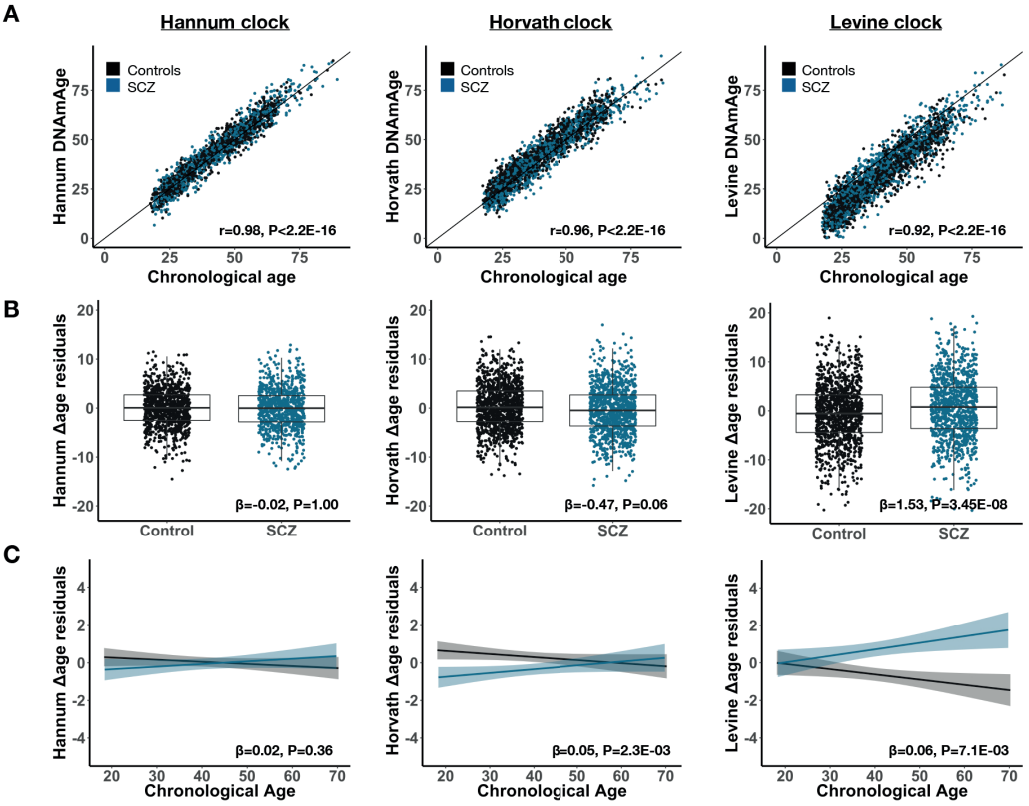


Figure 2. DNA methylation aging is altered in schizophrenia and conditional on chronological age. Presented are results visualizing DNAm aging in SCZ for each clock; Hannum (left), Horvath (middle), Levine (right). Cases are shown in blue and controls in black. (A) The correlation between DNAm age and chronological age. The Pearson's correlation estimate and corresponding p-value are shown in the bottom corner. (B) Boxplots of Δage between cases and controls with the meta-analytic effect size and p-value across cohorts shown. β represents the mean change in Δage in cases compared to controls. (C) Δage is visualized across chronological age with a regression line fitted separately for cases and controls and the meta-analytic interaction effect and p-value shown. β represents the change in Δage in cases per year of chronological age compared to controls. P-values are adjusted for multiple testing across clocks (n=3).

Modelling the interaction effect between disease status and chronological age on Δage reveals a differential rate of aging between cases and controls (Figure 2C). That is, the slope of Δage across chronological age is 0.05- and 0.06-years steeper in cases compared to controls for the Horvath (Pmeta=2.3E-03) and Levine clocks (Pmeta=7.1E-03), respectively, with no evidence of heterogeneity between cohorts (Figure S5 and Table S6). As no significant effects were observed for the Hannum Δage, we decided to focus our downstream analysis on the phenotypic (Levine) age and intrinsic cellular (Horvath) age only. To further disentangle the relationship between Δage in SCZ conditional on chronological age, we estimated differential aging by 10-year intervals, with years 18 and 19 included in the first age group. We observe significant DNAm age deceleration in early adulthood (18-30 years) with patients estimated at -1.23 years younger (Pmeta=3.9E-03) in intrinsic cellular age with no significant difference at later ages (Figure 3A). In phenotypic age, SCZ patients displayed significant DNAm age acceleration from 30 years and older (Figure 3B), with the most pronounced age acceleration between 50-60 years (2.29 years, Pmeta=9.0E-03). We again find no evidence of heterogeneity within age groups between cohorts (Figure S6 and Table S7-8).

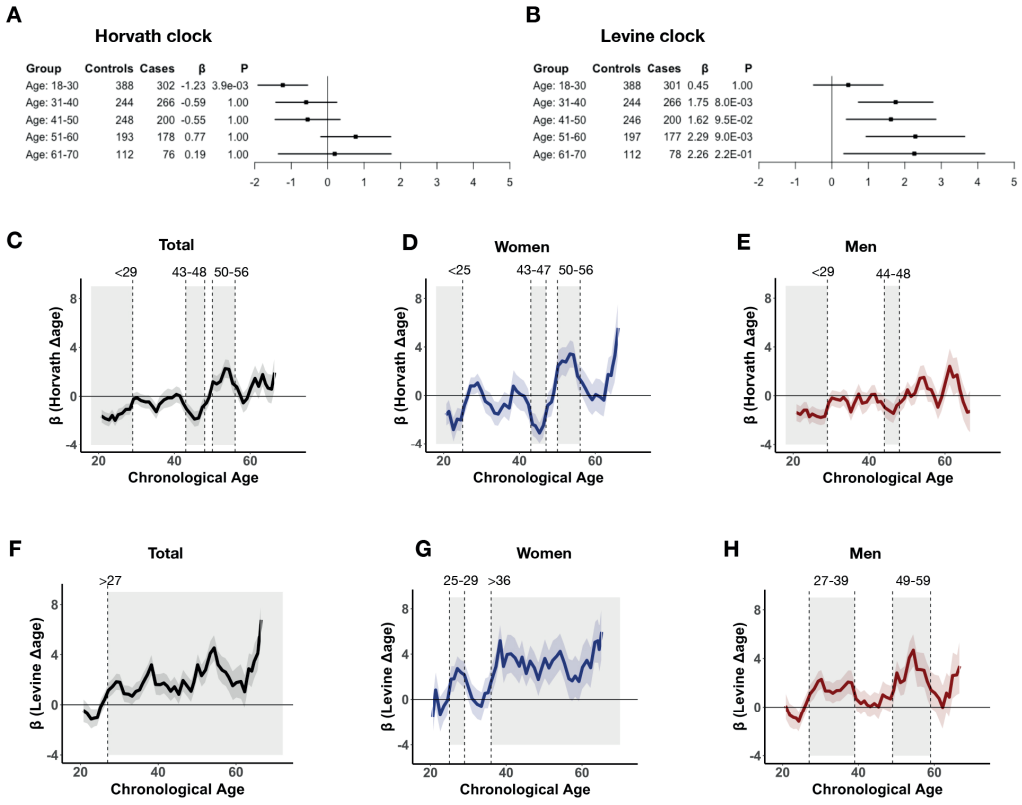


Figure 3. Differential DNAm aging in schizophrenia maps to specific age windows between sexes. *(A-B) Shown are Δage differences between cases and controls across age groups for the Horvath (A) and Levine clock (B). For each age group, number of cases and controls, and meta-analytic effect size (β) and p-value (P) are presented. P-values are corrected for multiple testing (2 clocks x 5 groups = 10 tests). See Table S5 for more details on results and*

corresponding statistics. (C-H) Sliding age-windows, using 5-year bins with steps of 1-year, were used to estimate differential aging (β) at finer resolution across the range of chronological age. Significant shifts in Δage between cases and controls, defined by the standard error of β deviating from zero for at least 3 steps, are highlighted by the shaded areas on the graph with the dotted vertical lines indicating the respective ages of the intervals. Identified age intervals for the Horvath and Levine clock are shown in C-E and F-H, respectively. Results for women (middle) and men (right) are presented in blue and red, respectively. The effects in the total sample are displayed in black (left).

Age- and sex-specific effects contribute to DNAm aging

To quantify the overall contribution of age- and also sex-specific effects, we estimated the gain in variance explained of Δage by adding the interaction terms of age and sex with disease status to a baseline model and assessed the gain in model performance. For both measures of aging, inclusion of interaction terms presented a significantly better fit, with the three-way interaction model (i.e. disease status, age and sex) explaining the most variance in Δage (Table 1 and S9). We observe a larger gain in model fit for the three-way interaction for phenotypic aging (P=0.01) than for intrinsic cellular aging (P=0.24), suggesting that sex-specific effects are more pronounced for Levine Δage.

		Horvath Δage		Levine Δage	
Model variables	**Model comparison**	**Δage R^2**	**P-value**	**Δage R^2**	**P-value**
Model 0: baseline	-	3.6%	-	2.1%	-
Model 1: + status	Model 0 vs 1	4.0%	6.6E-03	3.2%	4.4E-06
Model 2: + status*age.continuous	Model 1 vs 2	4.3%	0.08	3.7%	3.5E-03
Model 3: + status*age.groups	Model 1 vs 3	5.5%	1.4E-05	3.9%	0.02
Model 4: + status*age.groups*sex	Model 3 vs 4	5.9%	0.24	4.7%	0.01

Table 1. Age- and sex-specific effects significantly contribute to DNAm aging in schizophrenia. *Shown are the contributions of interaction effects between disease status and age and sex on Δage. The baseline model corresponds to Δage ~ dataset + ethnicity + platform + age.continuous + sex. For other models, the variable(s) in addition to the baseline variables are shown with the corresponding variance explained (RÇ) in Δage. Interaction terms with chronological age are modeled as a continuous variable (age.continuous) or a categorical variable (age.groups). The latter uses previously defined decades. Model comparison is performed to assess if the contribution of an interaction term is significant compared to a model without that term. The chi-square test is used to test two models with corresponding p-value presented. The results of these analysis are shown for both the Horvath and Levine clock. P-values are corrected for the number of tests performed (2 clocks x 4 comparisons = 8).*

Estimating and mapping windows of differential aging in schizophrenia

As our categorical age groups in the previous analyses were chosen somewhat arbitrarily, we conducted an exploratory analysis to refine age-dependent aging effects to identify specific age windows that are associated with differential aging. We implemented a sliding window approach across chronological age, both in the full sample and within each sex separately. Using 5-year

bins and sliding steps of 1 year, we tested cases versus age-matched controls and constructed a more precise picture of differential aging across chronological age in SCZ. At this finer resolution, we mapped changes in Δage to specific ages with different patterns between men and women. For intrinsic cellular age, we observe a deceleration effect during early adulthood from 29 years and younger across all samples, with the shift in differential aging occurring earlier in women (<25) (Figure 3C). For both men and women, we observe age deceleration in mid-forties and for women we also find age acceleration between 50-56 years (Figure 3C-E).

For phenotypic age, we mapped the age acceleration effect to 27 years and older across the whole sample with differences between the sexes (Figure 3F-H). In women, we find age acceleration between 25-29 years and from 36 years and older (Figure 3G). In men, we find age acceleration between 27-39 and 49-59 years (Figure 3H). More details on each age window and corresponding effect sizes are shown in Table S10. Thus far, our results show that DNAm aging, measured through the Horvath and Levine clock, is significantly different in SCZ and characterized by age-specific effects with some distinctions between the sexes, particularly for Levine Δage.

DNAm aging affects SCZ above and beyond smoking and blood cell types

To investigate the effect of smoking and blood cell type composition, we use DNAm-based smoking and cell type estimations (see Methods) as a proxy to evaluate their contribution to DNAm aging in SCZ. While DNAm clocks, by design, will encapsulate such effects, quantifying the contributions of each factor increases interpretability and helps understand the factors contributing to the differential aging findings. We observe that blood cell type proportions explain significantly more variance in DNAm aging than DNAm smoking scores (Supplementary Results S2.1). Inclusion of DNAm smoking score and blood cell proportions as covariates in our main models explains part but not all of the observed disease effects (Table S11 and Figure S8-9). Using a penalized regression framework (Table S12), we show that Levine Δage independently contributes to the variance in disease status in women older than 36 above and beyond smoking scores and blood cell type proportions (Supplementary Results S2.2 and Figure S10). A significant proportion of the Horvath Δage effect on disease status is reduced by adjusting for smoking (Table S11). However, smoking is not associated with Horvath Δage in controls (Pearson r=0.01, P=0.95) nor in cases (Pearson r=-0.08, P=0.28) (Figure S11). As smoking covaries with SCZ disease status, it is difficult to distinguish these signals.

Age deceleration by multi-tissue Horvath clock is not present in brain

We investigated DNAm aging in frontal cortex postmortem brain samples of 221 SCZ cases and 278 controls. The multi-tissue Horvath clock accurately predicts DNAm age in the brain as well (r=0.94, P < 2.2e-16). We, however, find no difference in DNAm aging between cases and controls (ß=-0.29, P=0.46) and no evidence of age-dependent aging. More details are shown in the Supplementary Results (S2.3).

Phenotypic age acceleration is associated with SCZ polygenic risk in women

To further decipher the factors underlying the signal of differential aging in SCZ, we examined the possible role of SCZ polygenic risk, age at onset, and illness duration (Figure S12). We first focus on the phenotypic age acceleration in female SCZ patients of age 36 years and older, as these individuals showed the most consistent and pronounced aging effect. We find stronger age

acceleration in cases with both low and high SCZ genetic risk (Table 2). More specifically, patients in the highest PRS1 tertile are predicted to be 4.30 years older in phenotypic age compared to controls (P=1.3E-05), patients with median range PRS1 are 1.89 years older (P=4.5E-02), and patients in the lowest quartile are 2.89 years older (P=2.8E-03). By permutation of PRS1 bins, we find that the effect in the highest PRS1 tertile is unlikely to occur by chance (P=0.024). For the association between Levine Δage and PRS1 to be most pronounced in the low and high tertile, is even less likely to happen by chance (P=0.006). At maximum, this group of women carrying high SCZ genetic risk have on average 7.03 higher phenotypic Δage (95% CI: 3.87-10.18; P=1.7E-05) (Figure 4A). We do not observe such an association in women age < 36 years, men with age > 36 years, nor across the whole dataset (Figure 4B and S13). Finally, by permuting the ranks of PRS1 within female cases >36 years, we find a mean maximum phenotypic Δage case-control difference of 3.69 years (95% CI: 1.26-6.12) across 1000 permutations, further demonstrating the significance of the observed maximum of +7.03 years phenotypic Δage difference. For age at onset and illness duration, we did not find significant association with Δage across partitioned bins (after permutation, P > 0.05) (Table 2). This is further confirmed when we integrated these two variables across PRS1 tertiles, demonstrating that the most pronounced differences in Δage are observed across PRS1 bins and not across the distribution of age at onset and illness duration in this subset of women (Figure S14).

Women: >36	Controls	Cases	Mean value in cases	β (Levine Δage)	95% CI	P
Polygenic risk						
All - no stratification	227	149	0.35	3.02	1.76 — 4.27	3.1E-06
PRS1 - continuous	-	149	0.35	0.42	-0.37 — 1.21	3E-01
PRS1 - low	227	50	-0.68	2.89	1.00 — 4.77	2.8E-03
PRS1 - mid	227	50	0.34	1.89	0.05 — 3.73	4.5E-02
PRS1 - high	227	49	1.40	4.30	2.40— 6.20	**1.3E-05***
Age of onset						
All - no stratification	227	111	26.50	3.73	2.34 — 5.12	2.3E-07
AOO - continuous	-	111	26.5	-0.08	-0.21 — 0.05	2.2E-01
AOO - early	227	37	17.43	3.26	1.11 — 5.41	3.1E-03
AOO - mid	227	37	25.43	3.70	1.58 — 5.81	6.7E-04
AOO - late	227	37	36.62	4.24	2.09 — 6.40	1.3E-04
Illness duration						
All - no stratification	227	111	23.37	3.73	2.34 — 5.12	2.3E-07
DUR - continuous	-	111	23.37	0.03	-0.07 — 0.13	6.1E-01
DUR - short	227	37	10.76	3.90	1.76 — 6.03	3.9E-04
DUR - mid	227	37	23.33	2.91	0.78 — 5.05	7.7E-03
DUR - long	227	37	36.01	4.39	2.25 — 6.53	7.3E-05

Table 2. Integration of Levine Δage with PRS, age of onset, and illness duration in women in later adulthood (precious page). *Analyses were performed using women >36 years of age. Only cases with available information were included in the analyses. Each phenotype was analyzed as both a continuous variable and as a categorical variable using equal tertiles from low to high bins. Mean values in cases for each phenotype are presented along with the association with Δage (β) and corresponding 95% confidence intervals and p-values. PRS1 = polygenic risk score PC1 (see Supplementary Information) scaled to mean zero with standard deviation of 1, AOO = age of onset, DUR = illness duration. Asterisk* indicates that significance (P < 0.05) by permutation analyses.*

We conducted a similar investigation on the observed intrinsic cellular age deceleration in all SCZ patients aged 29 years and younger but found no significant associations between Horvath Δage and PRS1, age at onset, or illness duration (Table S13 and Figure S15). While we did observe the strongest Horvath age deceleration in the high PRS1 tertile (β=-1.58, P=3.0E-03), this was not significant after permutation analysis (P>0.05). We did not analyse other identified age windows of differential aging as these either had too few individuals with genetic or phenotypic information available or more modest disease effects limiting any further stratification.

Finally, we assessed how Levine Δage and SCZ PRS1 compare in predicting SCZ disease status in our sample. Across the whole sample, PRS1 and Levine Δage explain 17.0% and 2.6% of the variance in disease status, respectively. Together, they explain 19.2%. In women in later adulthood, SCZ PRS1 and Levine Δage explain 11.5% and 9,8% independently and 22.4% jointly (Figure 4C).

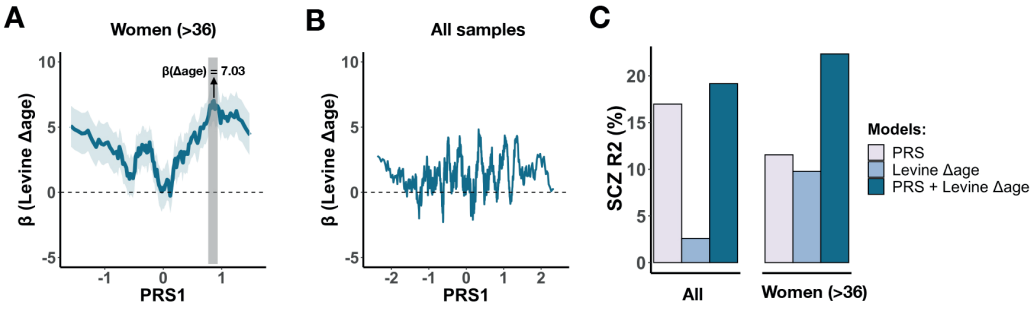


Figure 4. DNAm aging associates with SCZ PRS and additively contributes to SCZ disease status. *(A) Using a sliding-window approach, Levine Δage difference between cases and controls are shown across bins of ranked PRS1. Each bin contains 20 cases and slides from low to high PRS1 per shifts of one sample. The estimated Δage difference compared to all female controls >36 years is shown for each sliding bin in blue with the standard error in shaded blue. The most significant bin is highlighted by the grey vertical bar. (B) A similar analysis but then across all samples. (C) The variance explained in schizophrenia disease status (y-axis) by SCZ PRS and Levine Δage shown for all samples (left) and for women in later adulthood (right). The estimates shown are derived on top of the effect of sex, ethnicity, batch, platform, and chronological age.*

Discussion

We performed a large study of biological aging in schizophrenia using multiple epigenetic clocks based on whole blood DNA methylation data. We observe significant patterns of sex-specific and age-dependent DNAm aging in SCZ, a finding consistent across four European cohorts. The most significant differential aging pattern that we observe is in females ages 36 years and older in which we detect advanced phenotypic age acceleration, as measured by the Levine clock. We also observe intrinsic cellular age deceleration in SCZ patients during early adulthood, as measured by the Horvath clock. Phenotypic age acceleration in female patients is associated with a higher burden of SCZ polygenic risk. This high SCZ risk group displays accelerated aging of an average of +4.30 years compared to age-matched female controls. Phenotypic age and SCZ PRS furthermore contribute additively to SCZ and explain up to 22.4% of the variance in disease status. Our findings suggest that specific and identifiable patient groups are at increased mortality risk as measured by the Levine clock and warrant further research on DNAm clocks to examine its clinical relevance.

The Levine estimator was constructed by predicting a surrogate measure of phenotypic age, which is a weighted average of 10 clinical markers, including chronological age, albumin, creatinine, glucose and C-reactive protein levels, alkaline phosphatase and various blood cell related measures (Levine et al. 2018). By design, the Levine estimator is a composite biomarker that strongly predicts mortality, in particular that of age-related diseases, such as cardiovascular-related phenotypes. A 1-year increase in phenotypic age is associated with a 9% increased risk of all-cause mortality and a 10% and 20% increase of cardiovascular disease and diabetes mortality risk, respectively (Liu et al. 2018; Levine et al. 2018). Our findings of multiple year increase in phenotypic age in SCZ could thus imply an increased mortality in patients that is linked to cardiovascular disease, a previously well-established epidemiological observation (McGrath et al. 2008; Allebeck 1989b; Olfson et al. 2015). A recent study however found that DNAm age acceleration only predicts mortality in SCZ cases without pre-existing cancer using the Hannum clock (Kowalec et al. 2019). They did not find such evidence using the Levine clock. The smaller sample size and predominantly male cohort may have reduced the predictive power of the study. Our findings warrant a more focused and larger study of DNAm aging in female patients in later adulthood, preferably stratified by SCZ genetic risk. Our results align well with the observation that patients with SCZ, particularly women, are reported to be at high mortality risk due to cardiovascular disease and diabetes (Olfson et al. 2015; Osby et al. 2000a; Galletly et al. 2012). Assuming that cardiovascular risk is modifiable in SCZ (Kugathasan et al. 2018), phenotypic age could serve as a potential biomarker to identify at-risk individuals and in this way help with disease management and improvement of life-expectancy.

In contrast to age acceleration in phenotypic age, we observe age deceleration in intrinsic cellular age (i.e. the Horvath DNAm age), an effect that is most pronounced in patients age 29 and younger. Unlike the association findings in females, we did not observe clear patterns with genetic and phenotypic variables that could help to further decipher the signal. Horvath Δage furthermore showed strong age-specific effects but less clear sex-specific effects. We did not observe age deceleration in postmortem brain samples of the human cortex, indicating that the observed aging signal in SCZ may be blood-specific. Horvath DNAm aging has been shown to

be associated with molecular processes of development and cell differentiation (Horvath 2013; Horvath and Raj 2018), including through blood-based DNAm age measures in human (neuro) developmental phenotypes (Jeffries et al. 2019; Hoshino et al. 2019). Our findings may indicate that patients diagnosed with SCZ in this age group show evidence of delayed or deficient development and that this is detectable in blood through the multi-tissue Horvath clock. This however remains speculative and future work is needed to further dissect how blood-based Horvath age deceleration is associated with SCZ.

While we did observe aging effects with the Horvath and Levine clock, we did not with the Hannum clock. The Hannum clock is less predictive of age acceleration effects on mortality risk than the Levine clock (Levine et al. 2018), which could explain the lack of findings in our analyses. The Hannum estimator furthermore cannot be used on first generation 27K DNA methylation arrays which reduced the sample size of this study with 30% and may have impacted the statistical power of these specific analyses. This highlights the benefits of designing methods that are inclusive to all platforms, so all data, both old and new, can be leveraged.

After publication of the preprint of our manuscript (Ori et al., n.d.), Higging-Cheng et al. also reported significant DNAm alterations in SCZ (Higgins-Chen et al. 2020). This smaller study included 567 SCZ cases and 594 non-psychiatric controls with most of the sample (UK and SCT cohorts) also included in our study. Similar to our finding of 1.53 years of phenotypic age acceleration in schizophrenia cases, they report a 1.4- to 1.9-year increase in Δage in SCZ cases compared to controls. In addition, using GrimAge, a newly trained DNAm mortality clock (Lu et al. 2019), they observe age acceleration of 2.5- to 5.8-years. Unlike phenotypic age acceleration, this increase is largely driven by smoking effects. Similar to our work, this work highlights the value of analysing multiple clocks in conjunction and again suggesting that distinct biological processes of aging are altered in SCZ. In addition to the larger sample size, there are other key differences between our study and Higgins-Cheng et al. First, we performed detailed phenotypic analyses including explicit modelling of age and sex-specific effects. Second, methodically, we performed meta-analyses across cohorts as opposed to individual analyses per cohort. This approach, combined with multiple testing correction, is robust to cohort-specific artefacts in the data. Third, we integrated DNAm age with SCZ polygenic risk. Our PRS analyses yielded important insights into specific patient groups that could be at higher risk of all-cause mortality and that DNAm Δage and SCZ polygenic risk contribute additively to the illness. The latter suggests that combining genetic and epigenetic predictors can augment downstream prediction of outcomes in SCZ, similarly to what was recently shown for BMI (McCartney et al. 2018).

A systematic review of aging biomarkers found that less than a quarter of studies explored an interaction effect or statistically compared the regression slope between groups in SCZ (Nguyen, Eyler, and Jeste 2018). Our findings of sex-specific and age-dependent DNAm aging support their recommendations to specifically examine interaction effects with age and sex in aging studies but also more general in epigenetic studies of SCZ, such as epigenome-wide association studies. Future work should also be extended to integrate nonlinear models to fully capture the complex relationship between DNAm aging and clinically relevant variables across the lifespan of patients. These models will help validate and further refine the most relevant age intervals.

A limitation of the study is the cross-sectional design of the cohorts used. While we do find an association with SCZ polygenic risk, dissecting cause-and-effect relationships remains challenging. Independent replication studies are needed, preferably using longitudinal prospective cohorts with genomic data and information on symptom recurrence and severity, comorbidities and other phenotype-related variables. These studies can assess the clinical relevance of DNAm aging in SCZ above and beyond other known health risk factors and disease biomarkers, such as medication use. An urgent open question remains whether DNAm age signatures are modifiable with regards to clinical and lifestyle factors associated with SCZ. Improvement of existing methodology and/or development of new DNAm age biomarkers (Zhang et al. 2019; Bell et al. 2019) may in addition help to better study differential aging in SCZ and related disorders with increased mortality. Combining blood-based DNAm age with that of other aging profiles, such as MRI-based brain age (Schnack et al. 2016), may further advance our understanding of aging and SCZ disease progression, including the increased mortality (Cole et al. 2018). Finally, our findings support an integrative strategy with polygenic disease risk to improve clinical utilization.

Schizophrenia, like other mental illnesses, are associated with a wide-range of subsequent chronic physical conditions, including many age-related diseases (Scott et al. 2016). While health and life expectancy of the general population continues to improve, the mortality disparity between patients with schizophrenia and those unaffected continues to increase (Saha, Chant, and McGrath 2007; Hayes et al. 2017; Osby et al. 2000b; Lawrence, Hancock, and Kisely 2013). As the burden of age-related diseases continues to rise, early detection and subsequent opportunities for interventions before disabilities and co-morbidities become established will be important (Moffitt and Caspi 2019; Taylor and Reynolds 2020). Molecular biomarkers of aging, such as DNAm clocks, are now emerging as candidate tools for screening and intervention. Taken together, this study strengthens the need for more research on DNA methylation aging in SCZ, a population vulnerable to age-related diseases and excess mortality.

Declarations

Ethics approval and consent to participate

All cases and controls included in this study gave informed consent. Dutch (NLD) cohorts - ethical approval was provided by local ethics committees; University College London (UK) cohort - ethical approval was provided by National Health Service multicentre and local research ethics; Aberdeen (SCT) cohort - was provided by both local and multiregional academic ethical committees. Sweden (SWD) cohort - ethical permission was provided by the Karolinska Institutet Ethical Review Committee in Stockholm, Sweden.

Availability of data and materials

The datasets used are available on the NCBI Gene Expression Omnibus (GEO) data repository or through the principal investigator of each cohort. See Table S2 and S3 for an overview and corresponding accession series numbers. See Table S4 for sample information, including individual DNAm age estimates.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the US NIH under award number R01 DA028526, R01 MH078075, R21 MH098035, R01 MH115676, RF1AG058484 granted to RAO. LMOL was supported by the NIH under award number K99/R00 MH116115. PFS was supported by the Swedish Research Council (Vetenskapsrådet, award D0886501), the Horizon 2020 Program of the European Union (COSYN, RIA grant agreement n° 610307), and US NIMH (U01 MH109528 and R01 MH077139).

Authors' contributions

APSO and RAO conceived of the study. APSO performed data analyses and primary writing of the manuscript. LMOL, SH, and RAO advised on the work and co-wrote. RAO oversaw the work. RAO, RSK, EH, ED, DSC, NJB, AM, JM, JG and PFS provided access to data of cohorts included in the study. All authors read, gave input on, and approved the final manuscript.

Acknowledgements

We thank Dr. Hannah Elliott (University of Bristol MRC Integrative Epidemiology Unit) for providing code to calculate DNA methylation smoking scores. We thank all study participants for their participation in each of the respective cohorts.

References

Allebeck, P. 1989a. "Schizophrenia: A Life-Shortening Disease." Schizophrenia Bulletin.

Bell, Christopher G., Robert Lowe, Peter D. Adams, Andrea A. Baccarelli, Stephan Beck, Jordana T. Bell, Brock C. Christensen, et al. 2019. "DNA Methylation Aging Clocks: Challenges and Recommendations." Genome Biology 20 (1): 249.

Belsky, Daniel W., Avshalom Caspi, Renate Houts, Harvey J. Cohen, David L. Corcoran, Andrea Danese, Honalee Harrington, et al. 2015. "Quantification of Biological Aging in Young Adults." Proceedings of the National Academy of Sciences of the United States of America 112 (30): E4104–10.

Bergen, Sarah E., Alexander Ploner, Daniel Howrigan, CNV Analysis Group and the Schizophrenia Working Group of the Psychiatric Genomics Consortium, Michael C. O'Donovan, Jordan W. Smoller, Patrick F. Sullivan, Jonathan Sebat, Benjamin Neale, and Kenneth S. Kendler. 2019. "Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia." The American Journal of Psychiatry 176 (1): 29–35.

Boks, Marco P., Hans C. van Mierlo, Bart P. F. Rutten, Timothy R. D. J. Radstake, Lot De Witte, Elbert Geuze, Steve Horvath, et al. 2015. "Longitudinal Changes of Telomere Length and Epigenetic Age Related to Traumatic Stress and Post-Traumatic Stress Disorder." Psychoneuroendocrinology 51 (January): 506–12.

Carroll, Judith E., Michael R. Irwin, Morgan Levine, Teresa E. Seeman, Devin Absher, Themistocles Assimes, and Steve Horvath. 2017. "Epigenetic Aging and Immune Senescence in Women With Insomnia Symptoms: Findings From the Women's Health Initiative Study." Biological Psychiatry 81 (2): 136–44.

Chen, Brian H., Riccardo E. Marioni, Elena Colicino, Marjolein J. Peters, Cavin K. Ward-Caviness, Pei-Chien Tsai, Nicholas S. Roetker, et al. 2016. "DNA Methylation-Based Measures of Biological Age: Meta-Analysis Predicting Time to Death." Aging 8 (9): 1844–65.

Cole, James H., Riccardo E. Marioni, Sarah E. Harris, and Ian J. Deary. 2018. "Brain Age and Other Bodily 'Ages': Implications for Neuropsychiatry." Molecular Psychiatry, June.

Field, Adam E., Neil A. Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D. Adams. 2018. "DNA Methylation Clocks in Aging: Categories, Causes, and Consequences." Molecular Cell 71 (6): 882–95.

Fiorito, Giovanni, Silvia Polidoro, Pierre-Antoine Dugué, Mika Kivimaki, Erica Ponzi, Giuseppe Matullo, Simonetta Guarrera, et al. 2017. "Social Adversity and Epigenetic Aging: A Multi-Cohort Study on Socioeconomic Differences in Peripheral Blood DNA Methylation." Scientific Reports 7 (1): 16266.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software 33 (1): 1– 22.

Fries, Gabriel R., Isabelle E. Bauer, Giselli Scaini, Mon-Ju Wu, Iram F. Kazimi, Samira S. Valvassori, Giovana Zunta-Soares, Consuelo Walss-Bass, Jair C. Soares, and Joao Quevedo. 2017. "Accelerated Epigenetic Aging and Mitochondrial DNA Copy Number in Bipolar Disorder." Translational Psychiatry 7 (12): 1283.

Galletly, Cherrie A., Debra L. Foley, Anna Waterreus, Gerald F. Watts, David J. Castle, John J. McGrath, Andrew Mackinnon, and Vera A. Morgan. 2012. "Cardiometabolic Risk Factors in People with Psychotic Disorders: The Second Australian National Survey of Psychosis." The Australian and New Zealand Journal of Psychiatry 46 (8): 753–61.

Han, Laura K. M., Moji Aghajani, Shaunna L. Clark, Robin F. Chan, Mohammad W. Hattab, Andrey A. Shabalin, Min Zhao, et al. 2018. "Epigenetic Aging in Major Depressive Disorder." The American Journal of Psychiatry, April, appiajp201817060595.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." Molecular Cell 49 (2): 359–67.

Hayes, Joseph F., Louise Marston, Kate Walters, Michael B. King, and David P. J. Osborn. 2017. "Mortality Gap for People with Bipolar Disorder and Schizophrenia: UK-Based Cohort Study 2000-2014." The British Journal of Psychiatry: The Journal of Mental Science 211 (3): 175–81.

Higgins-Chen, Albert T., Marco P. Boks, Christiaan H. Vinkers, René S. Kahn, and Morgan E. Levine. 2020. "Schizophrenia and Epigenetic Aging Biomarkers: Increased Mortality, Reduced Cancer Risk, and Unique Clozapine Effects." Biological Psychiatry 88 (3): 224–35.

Hjorthøj, Carsten, Anne Emilie Stürup, John J. McGrath, and Merete Nordentoft. 2017. "Years of Potential Life Lost and Life Expectancy in Schizophrenia: A Systematic Review and Meta-Analysis." The Lancet. Psychiatry 4 (4): 295–301.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." Genome Biology 14 (10): R115.

Horvath, Steve, Wiebke Erhart, Mario Brosch, Ole Ammerpohl, Witigo von Schönfels, Markus Ahrens, Nils Heits, et al. 2014. "Obesity Accelerates Epigenetic Aging of Human Liver." Proceedings of the National Academy of Sciences, 201412759.

Horvath, Steve, and Kenneth Raj. 2018. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing." Nature Reviews. Genetics 19 (6): 371–84.

Horvath, Steve, and Beate R. Ritz. 2015. "Increased Epigenetic Age and Granulocyte Counts in the Blood of Parkinson's Disease Patients." Aging 7 (12): 1130–42.

Hoshino, Akina, Steve Horvath, Akshayalakshmi Sridhar, Alex Chitsazan, and Thomas A. Reh. 2019. "Synchrony and Asynchrony between an Epigenetic Clock and Developmental Timing." Scientific Reports 9 (1): 3770.

Jeffries, Aaron R., Reza Maroofian, Claire G. Salter, Barry A. Chioza, Harold E. Cross, Michael A. Patton, Emma Dempster, et al. 2019. "Growth Disrupting Mutations in Epigenetic Regulatory Molecules Are Associated with Abnormalities of Epigenetic Aging." Genome Research 29 (7): 1057–66.

Khoury, Louis El, Louis El Khoury, Tyler Gorrie-Stone, Melissa Smart, Amanda Hughes, Yanchun Bao, Alexandria Andrayas, et al. 2019 "Systematic underestimation of the epigenetic clock and age acceleration in older subjects" Genome Biology Dec 17;20(1):283.

Kirkpatrick, Brian, Erick Messias, Philip D. Harvey, Emilio Fernandez-Egea, and Christopher R. Bowie. 2008. "Is Schizophrenia a Syndrome of Accelerated Aging?" Schizophrenia Bulletin 34 (6): 1024–32.

Kowalec, Kaarina, Eilis Hannon, Georgina Mansell, Joe Burrage, Anil P. S. Ori, Roel A. Ophoff, Jonathan Mill, and Patrick F. Sullivan. 2019. "Methylation Age Acceleration Does Not Predict Mortality in Schizophrenia." Translational Psychiatry 9 (1): 157.

Kugathasan, Pirathiv, Henriette Thisted Horsdal, Jørgen Aagaard, Svend Eggert Jensen, Thomas Munk Laursen, and René Ernst Nielsen. 2018. "Association of Secondary Preventive Cardiovascular Treatment After Myocardial Infarction With Mortality Among Patients With Schizophrenia." JAMA Psychiatry.

Laursen, Thomas Munk, Merete Nordentoft, and Preben Bo Mortensen. 2014. "Excess Early Mortality in Schizophrenia." Annual Review of Clinical Psychology 10: 425–48.

Lawrence, David, Kirsten J. Hancock, and Stephen Kisely. 2013. "The Gap in Life Expectancy from Preventable Physical Illness in Psychiatric Patients in Western Australia: Retrospective Analysis of Population Based Registers." BMJ 346 (May): f2539.

Levine, Morgan E., H. Dean Hosgood, Brian Chen, Devin Absher, Themistocles Assimes, and Steve Horvath. 2015. "DNA Methylation Age of Blood Predicts Future Onset of Lung Cancer in the Women's Health Initiative." Aging 7 (9): 690–700.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." Aging 10 (4): 573–91.

Liu, Zuyun, Pei-Lun Kuo, Steve Horvath, Eileen Crimmins, Luigi Ferrucci, and Morgan Levine. 2018. "A New Aging Measure Captures Morbidity and Mortality Risk across Diverse Subpopulations from NHANES IV: A Cohort Study." PLoS Medicine 15 (12): e1002718.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, et al. 2019. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan." Aging 11 (2): 303–27.

Marioni, Riccardo E., Sonia Shah, Allan F. McRae, Brian H. Chen, Elena Colicino, Sarah E. Harris, Jude Gibson, et al. 2015. "DNA Methylation Age of Blood Predicts All-Cause Mortality in Later Life." Genome Biology 16 (1): 25.

McCartney, Daniel L., Robert F. Hillary, Anna J. Stevenson, Stuart J. Ritchie, Rosie M. Walker, Qian Zhang, Stewart W. Morris, et al. 2018. "Epigenetic Prediction of Complex Traits and Death." Genome Biology 19 (1): 136.

McGrath, John, Sukanta Saha, David Chant, and Joy Welham. 2008. "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality." Epidemiologic Reviews 30 (May): 67–76.

McKinney, Brandon C., Huang Lin, Ying Ding, David A. Lewis, and Robert A. Sweet. 2017. "DNA Methylation Evidence against the Accelerated Aging Hypothesis of Schizophrenia." NPJ Schizophrenia 3 (March): 13.

Moffitt, Terrie E., and Avshalom Caspi. 2019. "Psychiatry's Opportunity to Prevent the Rising Burden of Age-Related Disease." JAMA Psychiatry.

Nguyen, Tanya T., Lisa T. Eyler, and Dilip V. Jeste. 2018. "Systemic Biomarkers of Accelerated Aging in Schizophrenia: A Critical Review and Future Directions." Schizophrenia Bulletin 44 (2): 398–408.

Okazaki, Satoshi, Ikuo Otsuka, Shusuke Numata, Tadasu Horai, Kentaro Mouri, Shuken Boku, Tetsuro Ohmori, Ichiro Sora, and Akitoyo Hishimoto. 2019. "Epigenetic Clock Analysis of Blood Samples from Japanese Schizophrenia Patients." NPJ Schizophrenia 5 (1): 4.

Olfson, Mark, Tobias Gerhard, Cecilia Huang, Stephen Crystal, and T. Scott Stroup. 2015. "Premature Mortality Among Adults With Schizophrenia in the United States." JAMA Psychiatry 72 (12): 1172–81.

Ori, Anil P. S., Loes M. Olde Loohuis, Jerry Guintivano, Eilis Hannon, Emma Dempster, David St. Clair, Nick J. Bass, et al. n.d. "Schizophrenia Is Characterized by Age- and Sex-Specific Effects on Epigenetic Aging." https://doi.org/10.1101/727859.

Osby, U., N. Correia, L. Brandt, A. Ekbom, and P. Sparén. 2000a. "Time Trends in Schizophrenia Mortality in Stockholm County, Sweden: Cohort Study." BMJ 321 (7259): 483– 84.

Perna, Laura, Yan Zhang, Ute Mons, Bernd Holleczek, Kai-Uwe Saum, and Hermann Brenner. 2016. "Epigenetic Age Acceleration Predicts Cancer, Cardiovascular, and All-Cause Mortality in a German Case Cohort." Clinical Epigenetics 8 (June): 64.

Quach, Austin, Morgan E. Levine, Toshiko Tanaka, Ake T. Lu, Brian H. Chen, Luigi Ferrucci, Beate Ritz, et al. 2017. "Epigenetic Clock Analysis of Diet, Exercise, Education, and Lifestyle Factors." Aging 9 (2): 419–46.

Saha, Sukanta, David Chant, and John McGrath. 2007. "A Systematic Review of Mortality in Schizophrenia: Is the Differential Mortality Gap Worsening over Time?" Archives of General Psychiatry 64 (10): 1123–31.

Salomon, Joshua A., Juanita A. Haagsma, Adrian Davis, Charline Maertens de Noordhout, Suzanne Polinder, Arie H. Havelaar, Alessandro Cassini, et al. 2015. "Disability Weights for the Global Burden of Disease 2013 Study." The Lancet. Global Health 3 (11): e712–23.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." Nature 511 (7510): 421–27.

Schnack, Hugo G., Neeltje E. M. van Haren, Mireille Nieuwenhuis, Hilleke E. Hulshoff Pol, Wiepke Cahn, and René S. Kahn. 2016. "Accelerated Brain Aging in Schizophrenia: A Longitudinal Pattern Recognition Study." The American Journal of Psychiatry 173 (6): 607– 16.

Scott, Kate M., Carmen Lim, Ali Al-Hamzawi, Jordi Alonso, Ronny Bruffaerts, José Miguel Caldas-de-Almeida, Silvia Florescu, et al. 2016. "Association of Mental Disorders With Subsequent Chronic Physical Conditions: World Mental Health Surveys From 17 Countries." JAMA Psychiatry 73 (2): 150–58.

Sugden, Karen, Eilis J. Hannon, Louise Arseneault, Daniel W. Belsky, Jonathan M. Broadbent, David L. Corcoran, Robert J. Hancox, et al. 2019. "Establishing a Generalized Polyepigenetic Biomarker for Tobacco Smoking." Translational Psychiatry 9 (1): 92.

Taylor, Warren D., and Charles F. Reynolds. 2020. "Psychiatry's Obligation to Treat and Mitigate the Rising Burden of Age-Related Mental Disorders." JAMA Psychiatry.

Viana, Joana, Eilis Hannon, Emma Dempster, Ruth Pidsley, Ruby Macdonald, Olivia Knox, Helen Spiers, et al. 2017. "Schizophrenia-Associated Methylomic Variation: Molecular Signatures of Disease and Polygenic Risk Burden across Multiple Brain Regions." Human Molecular Genetics 26 (1): 210–25.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses inRwith themetaforPackage." Journal of Statistical Software 36 (3).

Voisey, Joanne, Bruce R. Lawford, C. Phillip Morris, Leesa F. Wockner, Ernest P. Noble, Ross Mcd Young, and Divya Mehta. 2017. "Epigenetic Analysis Confirms No Accelerated Brain Aging in Schizophrenia." NPJ Schizophrenia 3 (1): 26.

Zhang, Qian, Costanza L. Vallerga, Rosie M. Walker, Tian Lin, Anjali K. Henders, Grant W. Montgomery, Ji He, et al. 2019. "Improved Precision of Epigenetic Clock Estimates across Tissues and Its Implication for Biological Ageing." Genome Medicine 11 (1): 54.

Supplemental Materials

Full supplemental materials can be found here:
https://www.biorxiv.org/content/10.1101/727859v2.supplementary-material

S1. Supplementary Methods

S1.1 Cohort descriptions

The Netherlands (NLD)

The Dutch cohort is a case-control sample with inpatients and outpatients recruited from different psychiatric hospitals and institutions across the Netherlands, coordinated via academic hospitals in Amsterdam, Groningen, Maastricht, and Utrecht. Detailed medical and psychiatric histories were collected, including the Comprehensive Assessment of Symptoms and History (CASH), an instrument for assessing diagnosis and psychopathology. Only patients with a Diagnostic and Statistical Manual for Mental Disorders fourth edition (DSM-IV) diagnosis of schizophrenia were included as cases. All patients and controls were of Dutch descent, with at least three out of four grandparents of Dutch ancestry. The controls were volunteers and were free of any psychiatric history, the majority via the CASH. Whole blood DNA methylation data is collected for 586 cases and 516 controls. DNAm profiles were obtained using the Illumina's Infinium 27k Human DNA methylation Beadchip (v1.2) and the Infinium 450k Human DNA methylation Beadchip (v1.2). The 27k dataset can be found on GEO (GSE41037) with more details elsewhere (Horvath et al. 2012; van Eijk et al. 2015). Part of the 450k samples can be found under GSE41169. We have furthermore assayed an additional 200 controls and 196 cases.

University College London (UK)

More details on the cohort can be found elsewhere (International Schizophrenia Consortium 2008; Datta et al. 2010; Hannon et al. 2016). The UCL cohort is a case-control sample recruited from London and South England consisting of unrelated cases and ancestrally matched controls. All subjects were included if both parents were of English, Irish, Welsh or Scottish descent, with at least three out of four grandparents having the same origins. All cases were selected for having prior International Classification of Diseases 10 (ICD10) diagnosis of schizophrenia made by National Health Service (NHS) psychiatrists. The research subjects were then given interviews with the Schedule for Affective Disorders and Schizophrenia-Lifetime Version (SADS-L) schedule and further data were collected from NHS medical and nursing case notes and all other available sources. Therefore, all cases were selected on the basis of having a primary clinical diagnosis of schizophrenia made by a psychiatrist at interview according to ICD10 criteria and then at the probable level of schizophrenia with Research Diagnostic Criteria (RDC) made at interview by a second research psychiatrist. The control subjects were also interviewed with the initial clinical screening questions of the SADS-L and selected on the basis of not having a family history of schizophrenia, alcoholism or bipolar disorder and for having no past or present personal history of any RDC-defined mental disorder. Whole blood DNA methylation data (450K) is publicly available for 353 patients and 322 controls through GEO (GSE80417).

Aberdeen (SCT)

More details on the cohort can be found elsewhere (International Schizophrenia Consortium 2008; Datta et al. 2010; Hannon et al. 2016). The Aberdeen cohort is a case-control sample that contains patients diagnosed with schizophrenia and non-psychiatric controls who have self-identified as born in the British Isles (95% in Scotland). All cases met the DSM-IV and ICD-10 criteria for schizophrenia. Diagnosis was made by Operational Criteria Checklist (OPCRIT). Controls were volunteers recruited through general practices in Scotland. Practice lists were screened for potentially suitable volunteers by age and sex and by exclusion of individuals with major mental illness or use of neuroleptic medication. Volunteers who replied to a written invitation were interviewed using a short questionnaire to exclude major mental illness in the individual themselves and their first-degree relatives. Whole blood DNA methylation data (450K) is publicly available for 414 patients and 433 controls through GEO (GSE84727).

Sweden (SWD)

The Swedish cohort is a case-control sample that contains both patients and controls of older age, i.e. 50-70 years. Whole blood DNA methylation data (450K) is collected for 96 samples, for which after matching by predicted sex and genotype information, 37 cases and 32 controls were included in the analysis.

S1.2 Data preprocessing and normalization

IDAT files and thus raw fluorescence intensity values were available for the Dutch and Swedish cohorts while for the UK and Scottish cohorts methylated and unmethylated intensity values were downloaded from GEO data repository (Table S1). To analyze DNAm quantifications, we employed three data processing pipelines to accommodate the 450K and 27K arrays with raw data available and the 450K arrays with only methylation intensity values from GEO. For samples with IDAT files available, raw intensity values were read into an RGChannelSetExtended object in the R programming environment using the read.metharray() function in the minfi package (v1.20.2)(Aryee et al. 2014). For samples with IDAT files available (Dutch and Swedish cohort, n=1,212), we used probe detection P-values to exclude outlying samples. That is, samples with more than 5% of probes detected at $P > 0.05$ were excluded from further analyses (n=13). To capture technical variation, surrogate variables that cumulatively explained >90% of variation of the control probe intensity levels were estimated using the ctrlsva() function of the ENmix package (v1.10)(Xu et al. 2015) and stored to account for technical variance in downstream analyses. For the 450K arrays, background distributions were estimated using 600 chip internal control probes and dye-bias correction applied using the "RELIC" procedure(Xu et al. 2017) implemented through the preprocessENmix() function. Probe design type correction was applied by Regression on Correlated Probes (RCP) through the rcp() function of the ENmix framework(Niu, Xu, and Taylor 2016). For the 27K arrays, normal-exponential out-of-band (noob) correction, which applies background correction with dye-bias normalization, was applied using the preprocessNoob() function in minfi(Triche et al. 2013). For 450K arrays with only methylated and unmethylated intensity values available, background distributions were estimated and adjusted separately for each color channel and probe type using preprocessENmix() and probe design type correction subsequently applied using RCP.

S1.3 DNA methylation clocks

DNA methylation (DNAm), the addition of a methyl group to a cytosine nucleotide primarily at cytosine-phosphate-guanine (CpG) sites in the genome, is variable across the lifespan and has been shown to change relatively consistently between individuals(Bell et al. 2012; Christensen et al. 2009; Horvath et al. 2012). Variation across the epigenome can be aggregated and used to generate a highly accurate estimate of chronological age, known as DNAm age or the "epigenetic clock"(Horvath 2013; Hannum et al. 2013; Levine et al. 2018). The Hannum clock uses 71 probes with regression weights determined through a training dataset of whole blood 450K DNA methylation data in 656 adult samples, aged 19 to 101. The Hannum estimator accurately predicts age in whole blood samples of adults. The Horvath clock uses 353 CpGs, present on both the 27K and 450K array, with regression weights determined using 8,000 samples across 30 tissues and cell lines from children and adults across the lifespan, age 0 to 100 years (mean = 43, SD = 25). The Horvath estimator accurately predicts age across the lifespan and across tissues, including blood and brain. The Levine clock uses 513 CpGs, present on both the 27K and 450K array, with regression weights obtained by regressing the weighted average of ten routine clinical parameters, including age, on DNA methylation levels in whole blood of on average older adult samples. This generates an estimate of so called "phenotypic age" that is predictive of age in whole blood of older adults.

For each sample and each clock, DNAm age was estimated by incorporating DNAm levels of the pre-selected set of probes, identified by each estimator as predictive of chronological age, into a mathematical model that weighs each probe's methylation value and sums it to an aggregate DNAm age estimate. The Hannum and Horvath clock estimated DNAm age on average closer to chronological age (mean Δage Hannum = 1.0, Horvath = 1.7), while the Levine clock (i.e. phenotypic age) underestimated age by 7.7 years. See Table S1 for more details. The Levine clock was trained in an older adult population on a surrogate measure of biological age, generated through a Cox regression optimized to identify mortality-associated variables(Levine et al. 2018). It therefore underestimates chronological age at younger ages.

S1.4 Statistical analyses

For each clock and each cohort, we implemented a multivariable regression model predicting Δage as a function of schizophrenia status, sex, and age (model 1) and as well as predicting Δage as a function of sex and the interaction between schizophrenia status and chronological age (model 2). Chronological age and sex were included as covariates in all analyses, unless stated otherwise. Regression models were set up in R as follows;

model 1: lm(Δage ~ sex + age + status)
model 2: lm(Δage ~ sex + age + status + status:age)

For the Dutch cohort, batch and array platform were also included as covariates, as this cohort consists of multiple datasets from both the 27K and 450K platform. For each clock, regression coefficients with corresponding standard errors for each of the four cohorts were then supplied to the rma() function of the metafor package(Viechtbauer 2010) in R to fit a meta-

analytic fixed-effect model with inverse-variance weights and obtain an overall effect size and test statistic.

To quantify the significance of age- and sex-specific effects, we determined the contribution of interaction effects on top of the main disease effect. We first combined all cohorts to maintain necessary sample sizes across age and sex groups. Age categories were defined by grouping samples by decades with ages 18 and 19 included in the first decade (18-30, 31-40, etc.). To quantify the gain in variance explained in Δage, models with the interaction term were compared to a baseline model without the interaction term.

model 3: lm(Δage ~ cohort + sex + age.groups + status + **age.group:status**)
model 4: lm(Δage ~ cohort + sex + age.groups + status + age.groups:status + sex:status + age.group:sex + **age.group:sex:status**)

S1.5 Definition of age at onset across cohorts

The Netherlands

Age at onset is calculated preferentially by first using the date of first reported psychotic episode. If this information is not available the following order of variables is used as substitute; date of start treatment due to psychosis, date of start treatment with antipsychotic medication, date of first psychotic problems determined through the Comprehensive Assessment of Symptoms and History (CASH). Age at onset is available for 148 patients with overlapping DNAm data.

University College London (UK)

Age at onset is defined by the Operational Criteria Checklist (OPCRIT) and available for 321 patients.

Aberdeen (SCT):

Age at onset is available for 241 patients.

S1.6 PRS1 rationale and explanation

Polygenic risk scores (PRS) were generated by extracting each SNP's log odds ratio from an independent discovery GWAS of SCZ and applying these weights to SNPs in a second target dataset. For each individual, PRS is established by calculating the sum across weighted SNPs. As our cohorts were included in the discovery GWAS, the weights were derived from a leave-one-out analysis with that specific cohort excluded. Risk score profiles were calculated across ten GWAS P-value thresholds (5x10-8, 1x10-6, 1x10-4, 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, and 1.0), as previously described (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). While there is a rich correlation structure across p-value thresholds, the threshold with the most discriminatory power between cases and controls often differs between cohorts, which discourages from a one-size fits all approach. In addition, between-site variability and accompanying shifts in the distribution of polygenic risk can limit the interpretability of differential disease risk.

		NLD			SCT			UK		
PRS	Threshold	P	OR	ΔR2.prs	P	OR	ΔR2.prs	P	OR	ΔR2.prs
S1	5E-08	5,81E-10	1,44	4,03	2,96E-15	1,58	5,99	2,27E-09	1,51	4,89
S2	1E-06	1,58E-13	1,55	5,77	4,24E-21	1,76	8,75	1,07E-12	1,65	7,04
S3	1E-04	1,17E-18	1,73	8,49	5,74E-27	1,95	11,58	8,52E-24	2,21	14,93
S4	0,001	1,45E-23	1,88	11,05	2,01E-30	2,08	13,42	3,41E-29	2,54	19,04
S5	0,01	2,19E-28	2,06	13,69	2,41E-36	2,30	16,43	1,04E-29	2,61	19,53
S6	0,05	1,56E-29	2,13	14,48	3,71E-38	2,39	17,55	3,46E-31	2,75	20,98
S7	0,1	2,12E-29	2,13	14,42	2,02E-38	2,40	17,63	9,93E-32	2,86	21,69
S8	0,2	3,30E-29	2,12	14,29	5,96E-39	2,41	17,88	1,80E-30	2,75	20,52
S9	0,5	7,55E-29	2,11	14,10	4,38E-39	2,42	18,00	1,57E-31	2,85	21,47
S10	1	2,00E-28	2,09	13,87	9,07E-39	2,41	17,85	2,94E-31	2,83	21,25

Figure note: *each cohort with genetic data available has a different optimal GWAS p-value threshold that maximizes the magnitude of differential SCZ risk explained by PRS (highlighted in green). This demonstrates that choosing a single threshold across cohorts is suboptimal. NLD=Netherlands; SCT=Scotland, UK=United Kingdom, OR=odds ratio of PRS on disease status; ΔR2.prs=the unique variance explained in disease status by PRS.*

Recent work using principal component analysis (PCA) in combination with normalization of the PRS across sites has shown to effectively eliminate between-site variation and concentrate disease risk information into the first principal component (Bergen et al. 2019). This strategy works effectively because; (1) between-site variability can be eliminated by subtracting from all PRS values measured at one site the mean PRS among controls at that site, which aligns the PRS distribution with controls centered at zero; (2) PCA extracts disease relevant information across correlated thresholds of risk scores and concentrates this into a single component, which maximizes discriminatory power between cases and controls.

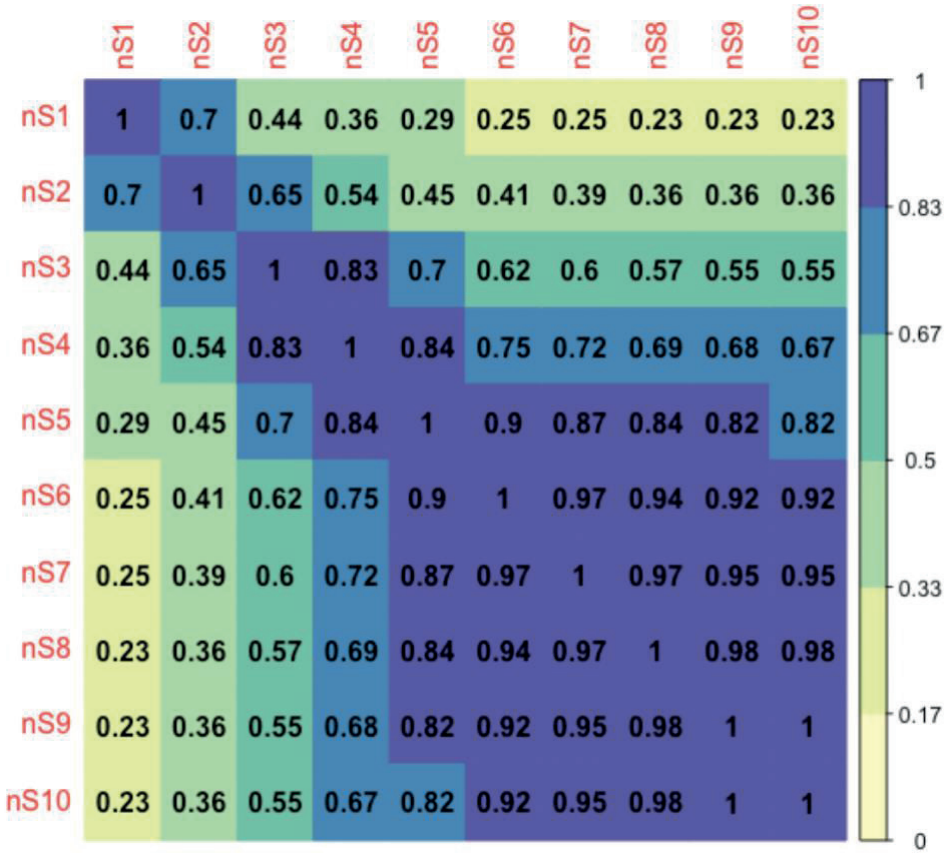


Figure note: shown is the correlation structure of SCZ PRSs computed across various GWAS p-value thresholds (S1-10). *While there is significant overlap between PRS of various thresholds, there is still unique information in each threshold that is not shared across threshold. The Pearson correlation is shown in each cell and color-coded as well. Principal component analysis can be used to extract disease risk information across thresholds and concentrating this in one dimension, represented by a single value.*

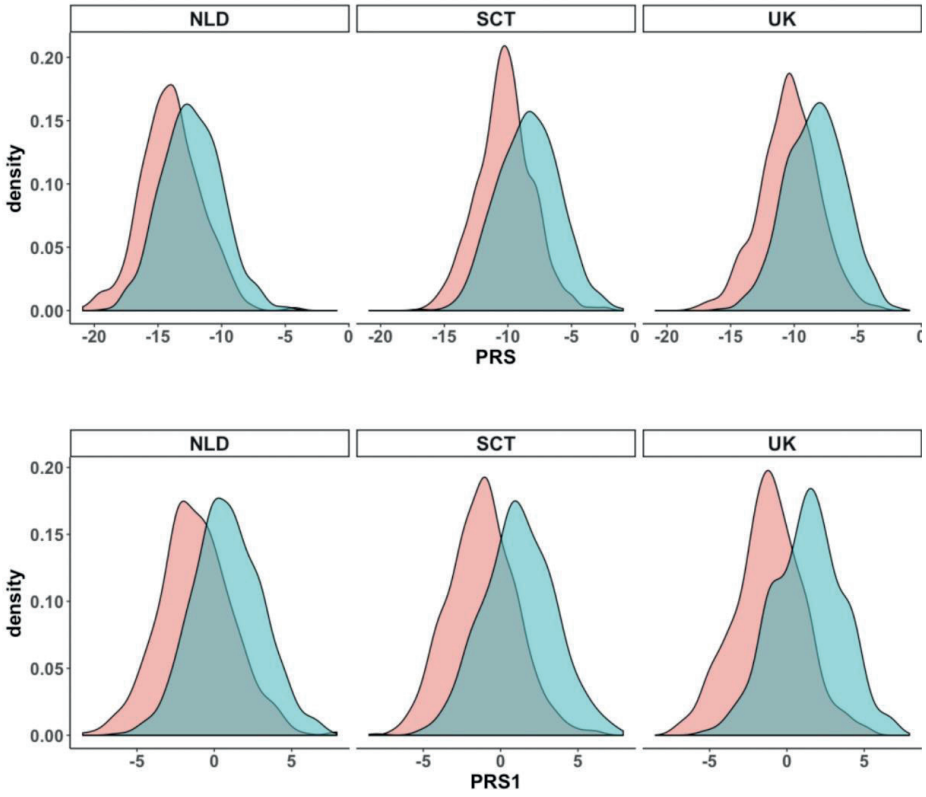
We therefore adopted the analysis framework of Bergen et al, 2019 and first aligned the PRS distribution for each cohort and p-value threshold at mean zero for controls. We then performed PCA on all the normalized scores, i.e. across thresholds and cohorts, to concentrated disease risk in the first principal component and refer to this single value as "PRS1". PRS1 explains 70.7% of the variance in PRS scores and captures 19.9% of the variance in disease status (OR=2.55) without adjusting for age and sex.

		Original scores				Normalized scores			
PRS	Threshold	P	OR	ΔR2.prs	ΔR2.site	P	OR	ΔR2.prs	ΔR2.site
S1	5E-08	1,09E-30	1,55	**4,90**	1,85	1,09E-30	1,50	**4,90**	0,67
S2	1E-06	3,27E-43	1,94	**7,12**	4,01	3,27E-43	1,64	**7,12**	0,58
S3	1E-04	6,78E-65	2,24	**11,22**	4,41	6,78E-65	1,91	**11,22**	0,51
S4	0,001	1,15E-78	2,33	**13,92**	3,64	1,15E-78	2,09	**13,92**	0,44
S5	0,01	1,47E-90	2,44	**16,28**	2,54	1,47E-90	2,26	**16,28**	0,28
S6	0,05	6,21E-95	2,87	**17,34**	5,26	6,21E-95	2,35	**17,34**	0,27
S7	0,1	2,85E-95	2,63	**17,43**	3,41	2,85E-95	2,36	**17,43**	0,25
S8	0,2	2,46E-94	2,73	**17,19**	4,02	2,46E-94	2,34	**17,19**	0,23
S9	0,5	6,04E-95	2,97	**17,37**	5,50	6,04E-95	2,36	**17,37**	0,23
S10	1	6,14E-94	2,78	**17,17**	4,36	6,14E-94	2,34	**17,17**	0,23

Figure note: *between-site variability can be eliminated by subtracting from all PRS values measured at one site the mean PRS among controls at that site. Importantly, while accounting for cohort differences, this normalization does not affect the variance explained by PRS overall compared to the original scores (see ΔR2.site between original and normalized scores). OR=odds ratio of PRS on disease status; ΔR2.prs=the unique variance explained in disease status by PRS; ΔR2.site=the unique variance explained in disease status by cohort.*

	Normalized scores					Principal components			
PRS	P	OR	ΔR2.prs	ΔR2.site	PCA	P	OR	ΔR2.prs	ΔR2.site
S1	1,09E-30	1,50	4,90	0,67	**PC1**	**1,90E-106**	**2,55**	**19,86**	**0,23**
S2	3,27E-43	1,64	7,12	0,58	PC2	1,7E-01	1,05	0,07	0,81
S3	6,78E-65	1,91	11,22	0,51	PC3	2,09E-01	1,04	0,06	0,79
S4	1,15E-78	2,09	13,92	0,44	PC4	9,75E-01	1,00	0,00	0,80
S5	1,47E-90	2,26	16,28	0,28	PC5	5,71E-01	1,02	0,01	0,81
S6	6,21E-95	2,35	17,34	0,27	PC6	4,62E-01	1,02	0,02	0,81
S7	**2,85E-95**	**2,36**	**17,43**	**0,25**	PC7	4,93E-01	1,02	0,02	0,80
S8	2,46E-94	2,34	17,19	0,23	PC8	4,88E-01	1,02	0,02	0,81
S9	6,04E-95	2,36	17,37	0,23	PC9	3,58E-01	1,03	0,03	0,80
S10	6,14E-94	2,34	17,17	0,23	PC10	1,35E-01	1,05	0,08	0,80

PCA on the normalized scores concentrates SCZ disease risk into the PC1 (highlighted in green). The discriminatory power by PC1 is superior to the original scores across various p-value thresholds. ΔR2.prs=the unique variance explained in disease status by PRS; ΔR2.site=the unique variance explained in disease status by cohort.



The distribution of PRS1 across cohorts (bottom panel) compared to PRS using the S6 P=0.05 threshold (top panel). PRS1 showns greater discriminatory power between cases and controls for each of the three cohorts.

S1.7 DNAm-based estimation of smoking

We computed a DNAm-based smoking score using CpG sites significantly associated with smoking. This smoking estimate has been shown to be a good proxy for smoking status (Elliott et al. 2014) and successfully used to account for the effects of smoking in large-scale DNAm studies of schizophrenia (Hannon et al. 2016). In a similar fashion as these studies, we calculate a weighted score across 183 DNA methylation sites, with the weights being effect sizes obtained from an epigenome-wide association study of smoking (Zeilinger et al. 2013).

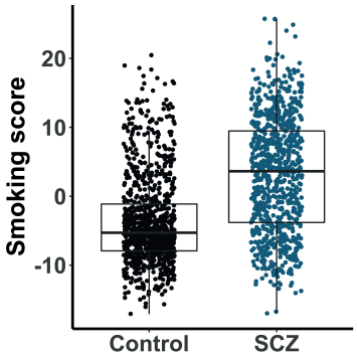
S1.8 DNAm-based estimation of blood cell types

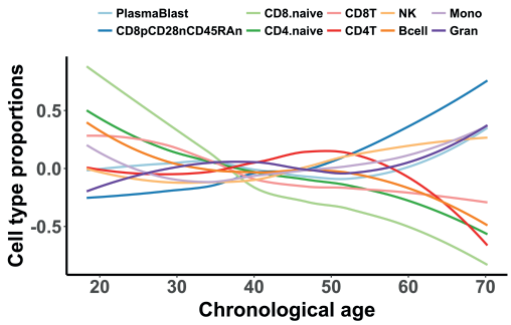
Estimated blood cell-type proportions were computed for all 450K array samples using the Horvath Methylation Age Calculator software. Further details on the derivation of estimates on CD8 T cells, CD4 T cells, nature killer cells, B cells, monocytes, and granulocytes can be found here (Houseman et al. 2012) and for plasma blasts, CD8+CD28-CD45RA-T cells, naive CD8 T cells here (Horvath 2013).

S2. Supplementary Results

S2.1 Analysis of DNAm aging in relation to estimates of smoking and blood cell types

Patients with SCZ smoke more than the general population(Kelly and McCreadie 2000) and blood cell type composition changes across the lifespan(Jaffe and Irizarry 2014). To investigate the effect of these factors, we use DNAm-based smoking and cell type estimations (see Methods) as a proxy to evaluate their contribution to DNAm aging in SCZ.





While DNAm clocks, by design, will encapsulate such effects, quantifying the contributions of each factor increases interpretability and helps understand the factors contributing to the differential aging findings. Here, we quantified the effects of DNAm smoking scores and blood cell types proportions on Horvath and Levine Δage in relation to the effect of disease status. For the Horvath clock, a baseline model with batch, ethnicity, sex, and age (as continuous) explains 3.9% of the variance in Δage. The addition of DNAm smoking scores and blood cell type proportions increases the fit of the model to 4.9% and 8.2%, respectively. The baseline model with both smoking scores and cell type estimates explain 9.4% of the variance in aging. While this reduces the main effect of disease status, the interaction between status and age (as categorical variable) remains significant and thus, as far as we can measure, independent of smoking and cell type composition.

		Horvath Δage		Levine Δage	
Model variables	**Model comparison**	**Δage R^2**	**P-value**	**Δage R^2**	**P-value**
Model x: baseline		3.9%	-	3.1%	-
Model y: baseline (+ smoking)	-	4.9%	-	5.3%	-
Model z: baseline (+ cell types)	-	8.2%	-	22.1%	-
Model 0: baseline (+ smoking/cell)	-	9.4%	-	22.8%	-
Model 1: + status	Model 0 vs 1	9.3%	0.86	22.8%	0.26
Model 2: + status*age.continuous	Model 1 vs 2	9.4%	0.10	23.2%	1.3E-03
Model 3: + status*age.groups	Model 1 vs 3	10.6%	2.1E-04	23.2%	0.05
Model 4: + status*age.groups*sex	Model 3 vs 4	10.8%	0.15	23.6%	0.05

Table S11. Age- and sex-specific on DNAm aging in schizophrenia adjusted for smoking and cell type estimates. *Shown are the contribution of interaction effects between disease status and age and sex on Δage when adjusted for DNAm smoking scores (baseline model y) and blood cell type proportions (baseline model z). The full baseline model is defined as Δage ~ dataset + ethnicity + age.continuous + sex + DNAm smoking score + DNAm blood cell type proportions. For other models, the variable(s) in addition to the full baseline variables are shown with the corresponding variance explained (R^2) in Δage. Interaction terms with chronological age are modelled as a continuous variable (age.continuous) or a categorical variable (age.groups). The latter uses previously defined decades. Model comparison is performed to assess if the contribution of an interaction term is significant compared to a model without that term. The chi-square test is used to test two models with corresponding p-value presented. The results of these analysis are shown for both the Horvath and Levine clock. These analyses included only 450K samples for which smoking scores and cell type estimates can be computed (N=1,621, 867 controls, 754 cases).*

For the Levine clock, a baseline model with batch, ethnicity, sex, and age (as continuous) explains 3.1% of the variance in Δage. The addition of DNAm smoking scores and blood cell type proportions increases the fit of the model to 5.3% and 22.1%, respectively. The baseline model with both smoking scores and cell type estimates explain 22.8% of the variance in aging. This indicates that blood cell type composition explains a large proportion of the variance in Levine Δage, which is expected as the Levine clock is trained on blood mortality markers, including blood cell counts. Modeling smoking scores and blood cell type proportions as covariates reduces the main effect of disease status on Levine Δage. The interaction between status and age (as continuous variable) remains significant, similar to the analysis of Horvath Δage. For Levine Δage, the three-way interaction of status, age, and sex shows a slightly better model fit, after accounting for smoking and cell types, as well.

S2.2 Levine Δage affects schizophrenia independently from smoking and cell types in women

In women age 36 and older, the patient group in which we observed the most profound aging effects, a lasso logistic regression selected dataset/ethnicity, smoking, 5 cell types, and Levine Δage as independent variables to explain a total of 23.6% of the variance in SCZ disease status (P=2.2E-08) (Table S12). Levine Δage explains 7.7% individually and 2.8% (P=3.3E-03) when adjusted for other selected variables (Figure 6C).

In individuals 29 years and younger, the group with significant Horvath deceleration aging, the lasso regression selected batch/ethnicity, age, sex, smoking, 7 cell types, and Horvath Δage as independent variabls to explain 49.8% of the variance in disease status (Table S12). A large proportion of this effect is driven by smoking, which explains 28.8%. Horvath Δage explains 3.1% of the variance in SCZ individually and 0.6% adjusted for other select variables (P=0.14). A significant proportion of the Horvath Δage effect on disease status is reduced by adjusting for smoking. However, smoking has no association with Horvath Δage in controls (Pearson r=0.01, P=0.95) nor in cases (Pearson r=-0.08, P=0.28) (Figure S12). As smoking covaries with SCZ disease status, it is difficult to distinguish these signals. In relation to SCZ genetic risk, smoking and blood cell types demonstrate limited effects on the observed pattern of differential aging across PRS1 (Figure S13).

S2.3 Analysis of postmortem brain samples

DNAm age estimates were generated for postmortem brain frontal cortex samples using publicly available data across four datasets (Table S2). The same data processing steps as described above were used for sample filtering and to generate beta values as input to estimate DNAm age. We excluded 16 samples due to missing data or discrepancy between reported and predicted sex and included only adult (age >= 18) frontal cortex samples, leaving 499 samples, 221 cases and 278 controls. Given the available sample sizes for some of the cohorts, we modeled Δage as a function of age and sex and disease status while correcting for cohort instead of performing a meta-analysis.

Dataset	Ancestry	Platform	Tissue	Data type	Total	Cases	Controls	Age (sd)
GSE74193	AA/EUR	450K	DLPFC	IDAT files	503	224	279	46.9 (15.4)
GSE61107	-	450K	Frontal cortex	IDAT files	48	24	24	61.7 (19.2)
GSE61380	-	450K	Frontal cortex	(un)methylated intensities	33	18	15	44.0 (15.7)
GSE61431	-	450K	Frontal cortex	(un)methylated intensities	43	20	23	61.8 (17.5)
Total	Mixed / unknown	450K	Frontal cortex	Mixed	627	286	341	

Figure note: *multiple datasets of postmortem brain DNAm data were included in in the analysis. Shown above are some sample characteristics and accompanying GEO accession numbers for each dataset before quality control. AA = African American, EUR = European, DLPFC = Dorsolateral prefrontal cortex.*

Only the Horvath clock yielded DNAm age estimates that closely correlated with chronological age. While the Hannum and Levine clock demonstrated decent correlations as well, they significantly underestimated chronological age. We therefore only investigated the Horvath clock, a multi-tissue estimator, and analyzed differential aging between cases and controls.

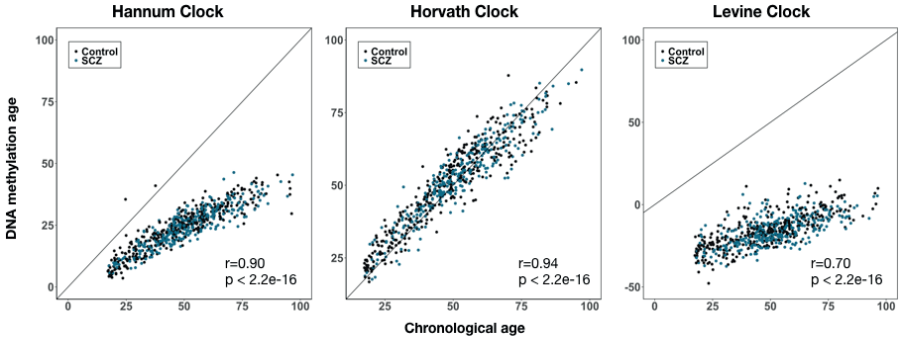


Figure note: *correlation plots between DNAm age estimates and chronological age for postmortem brain samples. Results are shown for the Hannum (left), Horvath (middle), and Levine clock (right). Cases are colored blue and controls black. Pearson correlation and corresponding p-values are shown in the right bottom corner.*

As with our analyses in blood, we first removed samples with descrepancies between reported sex and DNAm-based estimated sex (n=14) and samples with missing age information (n=2). Next, we regressed Δage on principal components of the control probes that explain >90% of the variation in control probe intensity values. The residuals were then added on mean(Δage) to generate Δage-adjusted, which preserves Δage in interpretable units (i.e. years). Using a multivariable linear regression model, we estimated differential DNAm aging in SCZ as follows;

model 1: lm(Δage ~ dataset + sex + age + status)
model 2: lm(Δage ~ dataset + sex + age + status + status:age)

Across the full dataset, we found no difference between cases and controls (ß=-0.29, P=0.46). We in addition did not observe a significant age-dependent disease effect, neither modeling age as a continuous variable (P=0.20) nor as a categorical variable (P=0.11). We also did not find Horvath age deceleration during early adulthood (age 18-30, ß=-0.09, P=0.92), like we observed in blood. We therefore conclude that there is no differential DNAm aging in the frontal cortex using postmortem brain samples.

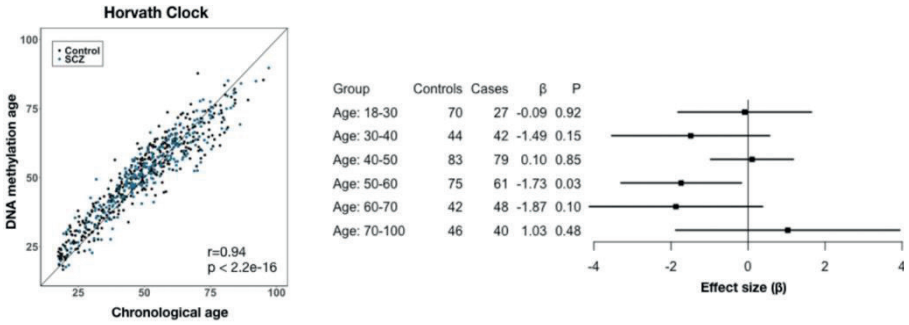


Figure note: *results of Horvath DNAm aging in frontal cortex postmortem brain samples of SCZ cases and controls. The left plot shows the correlation between DNAm age estimates and chronological age. The right plot shows a forest plot of the aging effects (ß) between cases and controls across various age groups.*

Supplementary Figures

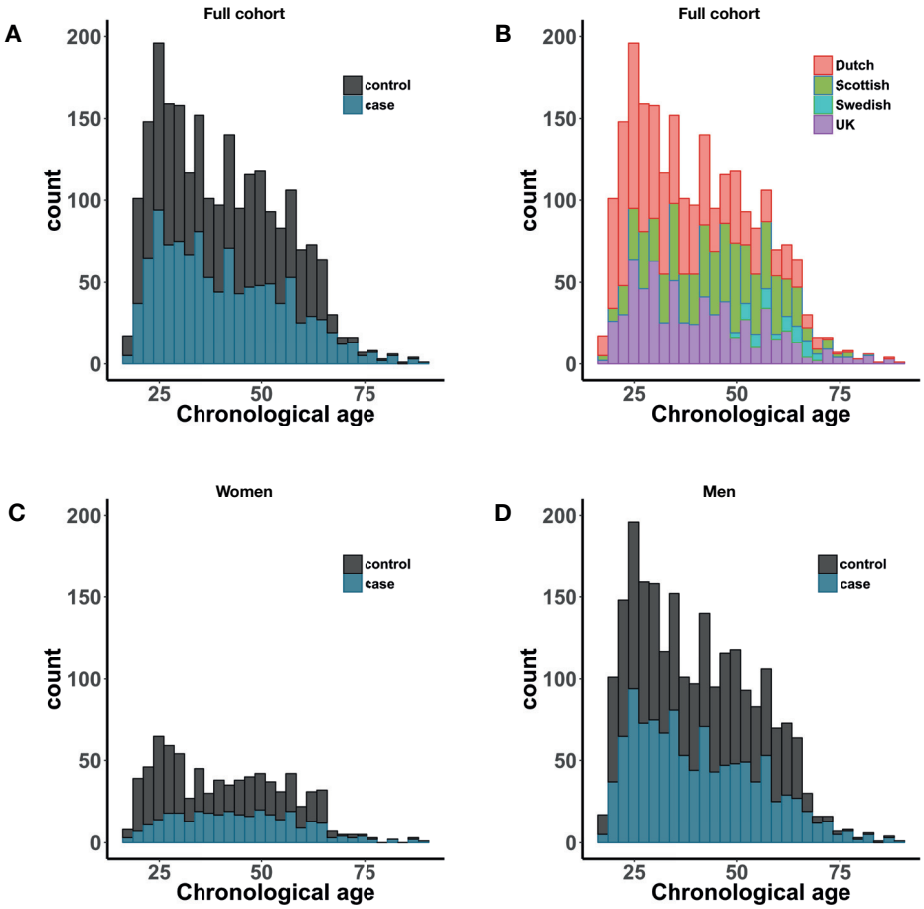


Figure S1. Sample distribution across chronological age. *(A) Full sample colored by disease status. (B) Full sample colored by ethnicity. Women (C) and men (D) colored by disease status.*

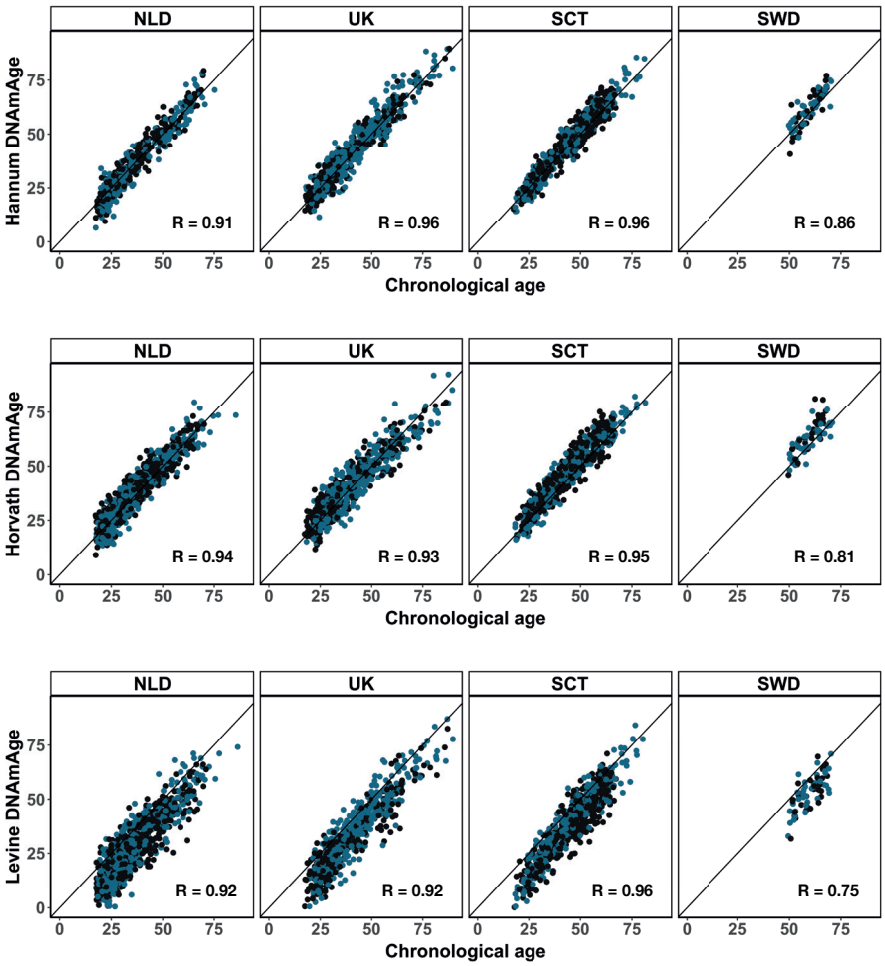
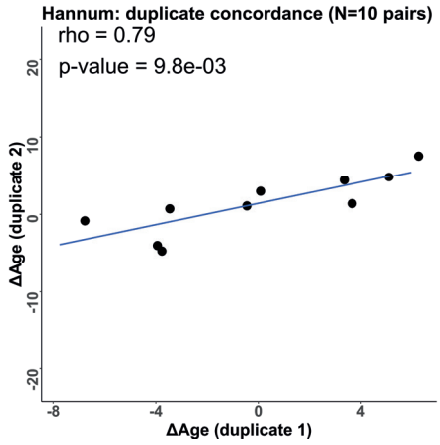
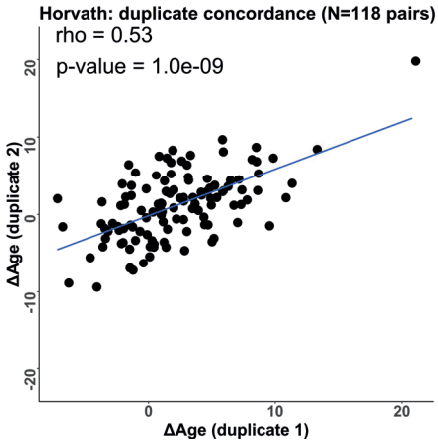


Figure S2. Correlations between DNAm age and chronological age by ethnicity. *Shown are correlations for the Hannum (top), Horvath (middle), and Levine clock (bottom) across cohorts.*





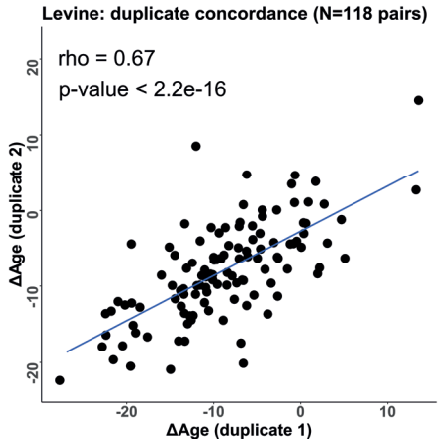


Figure S3. Sample duplicate pair concordance for DNAm age estimates of the Horvath and Levine clock. *Using Δage across 118 duplicate pairs, the concordance between pairs is shown for the Horvath (top-left) and Levine (top-right) plot. For the Hannum clock (bottom), only duplicates with both samples on the 450K array (N=10) could be used.*

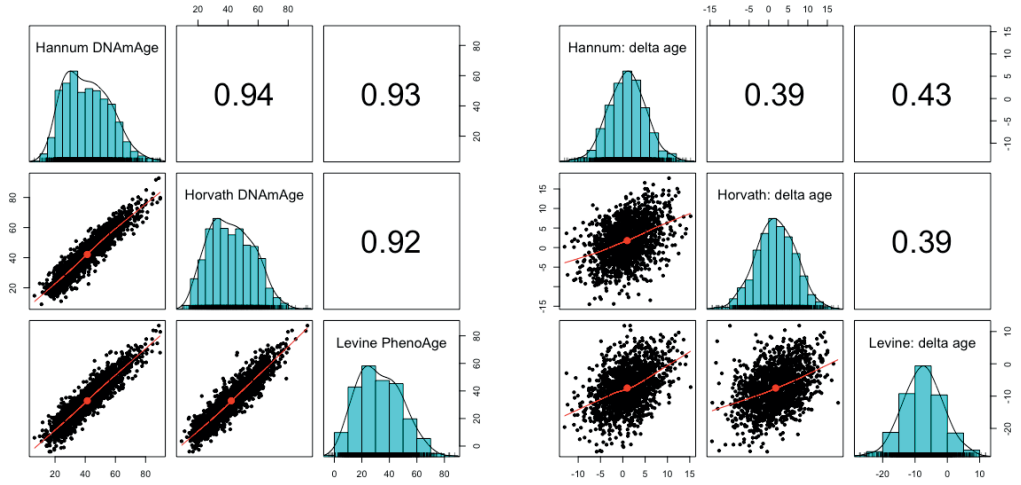


Figure S4. Correlation structure across clocks highlights that they capture both shared and distinct aspects of aging. *Shown are pair-wise scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation above the diagonal for DNAm age (left) and Δage (right) across the three clocks.*

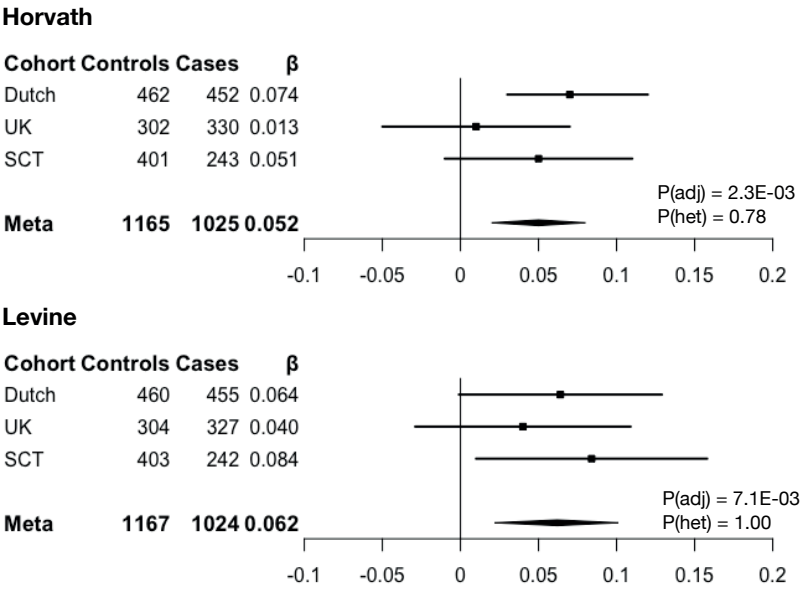


Figure S5. Differential DNAm aging across cohorts. *Shown are results for modeling the interaction between disease status and chronological to estimate Δage differences between cases and controls conditional on age. For each estimator - Hannum (top), Horvath (middle), Levine (bottom) - number of cases and controls, and meta-analytic effect size (β) and adjusted p-value (Padj) are presented. See Table S4 for more details on results and corresponding statistics. The Swedish cohort was excluded from this analysis as it has only a limited spread in age (i.e. 50-70 years).*

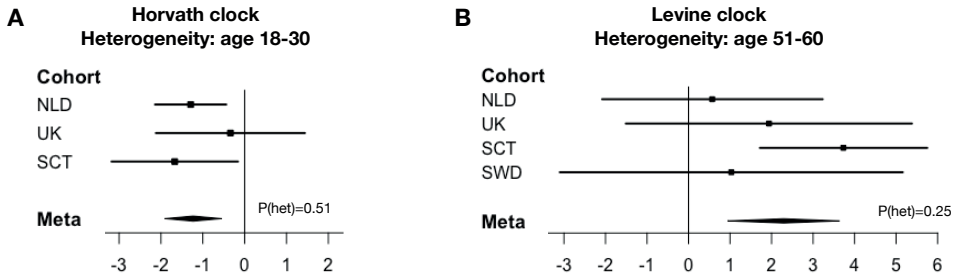


Figure S6. Heterogeneity of aging effects within age groups across cohorts. *Shown are results for Δage differences between cases and controls in specific age groups for the Horvath (left) and Levine (right) clock. Each forest plots show a significant meta-analytic effect size and the p-value of Cochran's heterogeneity test (Phet). See Table S7 and S8 for more details on results and corresponding statistics. The Swedish cohort was excluded from the left plot as it has only a limited spread in age (i.e. 50-70 years).*

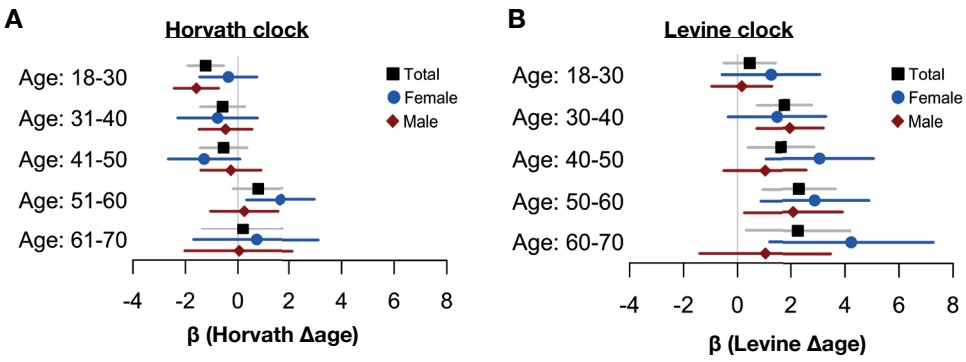


Figure S7. Sex-stratified differential aging by age groups in schizophrenia. *Shown are Δage differences between cases and controls across age groups stratified by sex for the Horvath (A) and Levine clock (B). Results for women and men are presented in blue and red, respectively. The effects in the total sample are displayed in black.*

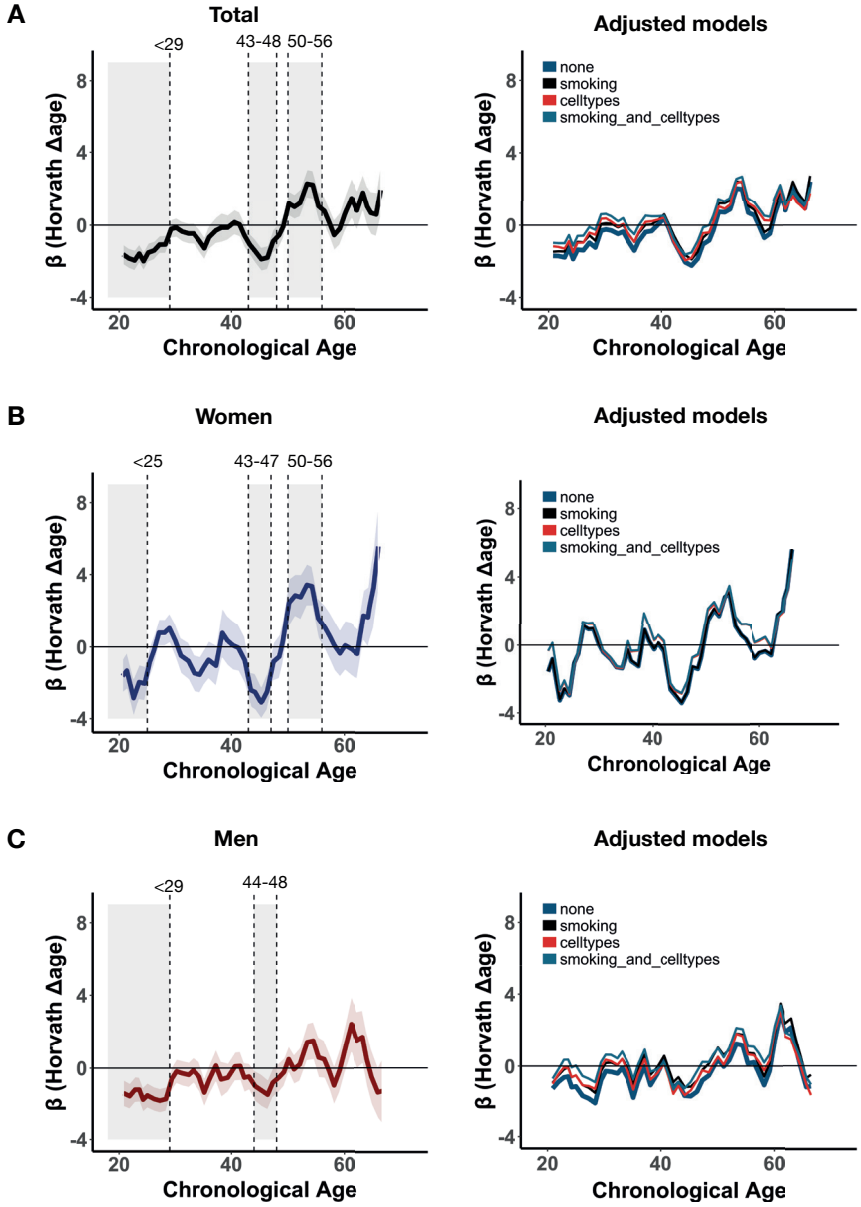


Figure S8. Horvath differential aging across chronological age adjusted for DNAm smoking and cell type proportions. Sliding age-windows, using 5-year bins with steps of 1-year, were used to estimate differential aging (β) at finer resolution across the range of chronological age. Graphs on the left show results without adjustment for smoking and cell types. Graphs on the right show results with adjustment for smoking and cell types. A) total sample size, B) women only, C) men only. For the right graphs only 450K samples were included, as DNAm smoking and cell types estimates cannot be calculated for 27K samples. These analyses therefore include less samples.

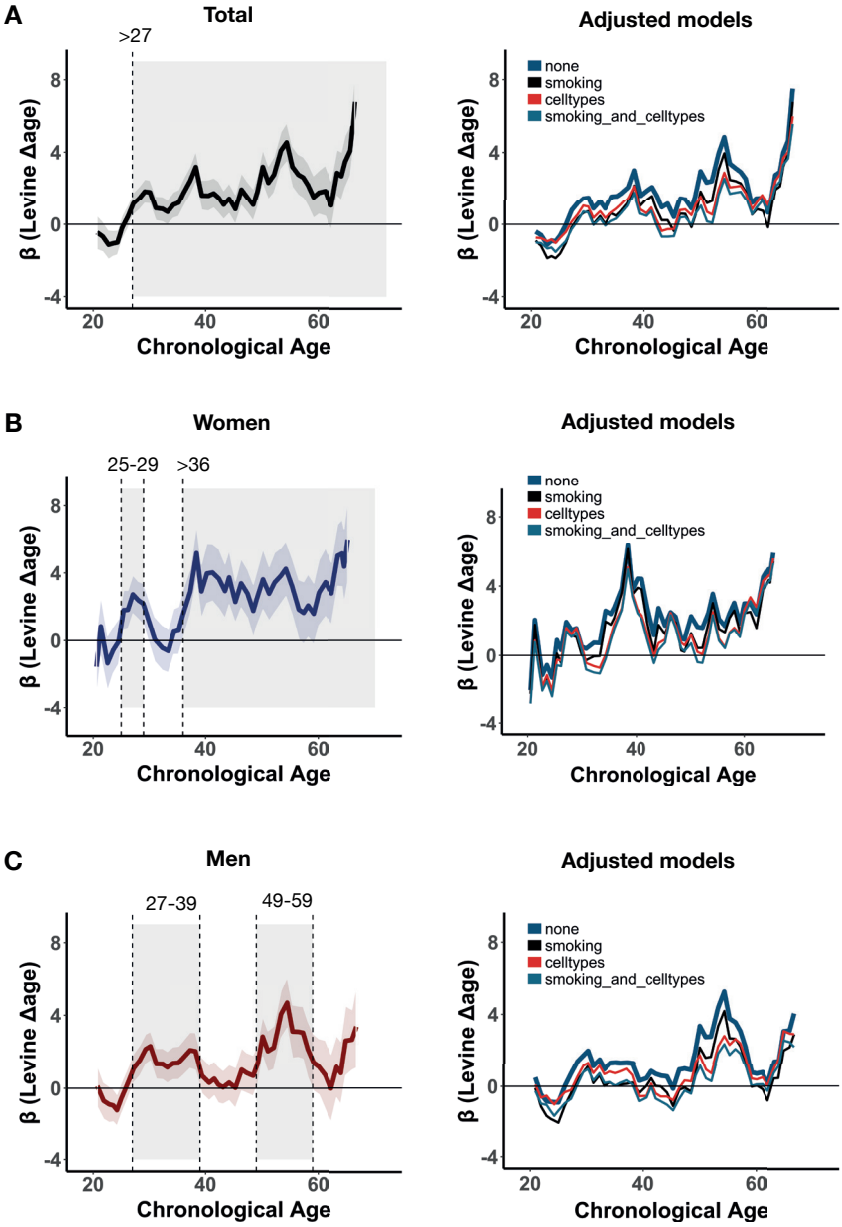


Figure S9. Levine differential aging across chronological age adjusted for DNAm smoking and cell type proportions. Sliding age-windows, using 5-year bins with steps of 1-year, were used to estimate differential aging (β) at finer resolution across the range of chronological age. Graphs on the left show results without adjustment for smoking and cell types. Graphs on the right show results with adjustment for smoking and cell types. A) total sample size, B) women only, C) men only. For the right graphs only 450K samples were included, as DNAm smoking and cell types estimates cannot be calculated for 27K samples. These analyses therefore include less samples.

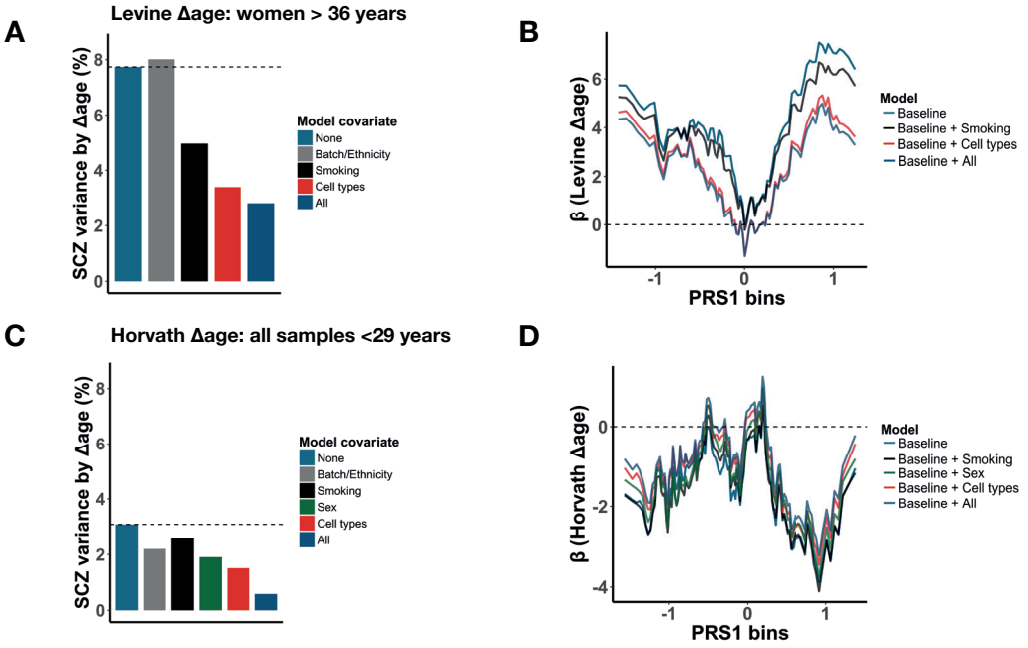


Figure S10. Smoking and blood cell type composition contribute in part to DNAm aging. *Presented are the results of a sensitivity analysis of DNAm-based estimated smoking score and blood cell type proportions in the 450K subsample of the cohort. (A/C) The proportion of schizophrenia variance explained by Δage after adjustment of various variables that are significantly associated with disease status. The "All" model presents the variance explained by Δage independent from all other variables. (B/D) Δage effect size is shown across bins (N = 20 cases/bin) of ranked PRS1 (unit = SD) for several models that adjust for covariates. The baseline model represents the effect of Δage adjusted for batch, ethnicity and chronological age. Results are shown for Levine Δage women > 36 years, 99 cases and 181 controls (C) and Horvath Δage all samples < 29 years, 141 cases and 238 controls (D) separately.*

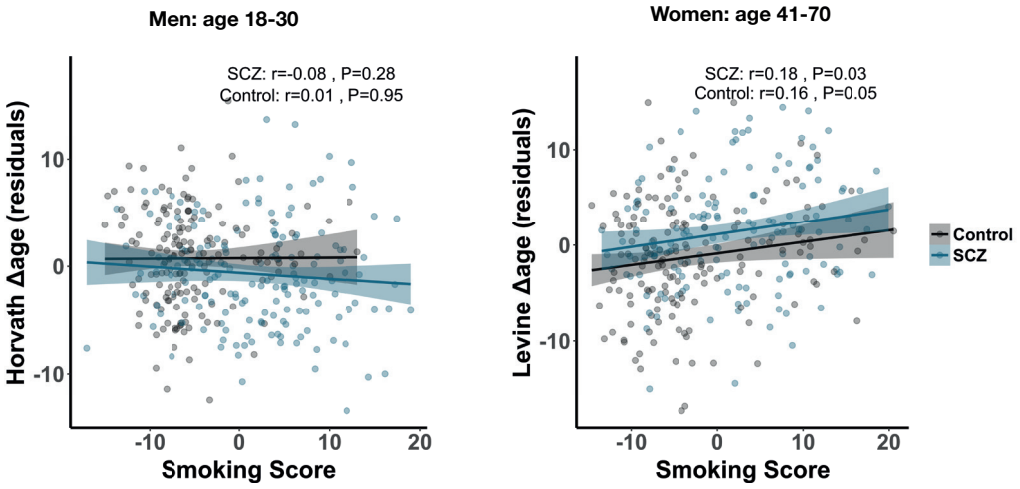


Figure S11. Smoking has only minimal effects on DNAm aging. *Shown is the relationship between Δage and DNAm-based smoking score for men (left plot; age 18-30, case = 165, control = 163) and women (right plot; age 41-70, case=144, control=159). DNAm Δage was first regressed on batch, ethnicity and chronological age. The Pearson correlations and corresponding p-values are shown on top of each plot. Cases and controls are presented in blue and black, respectively.*

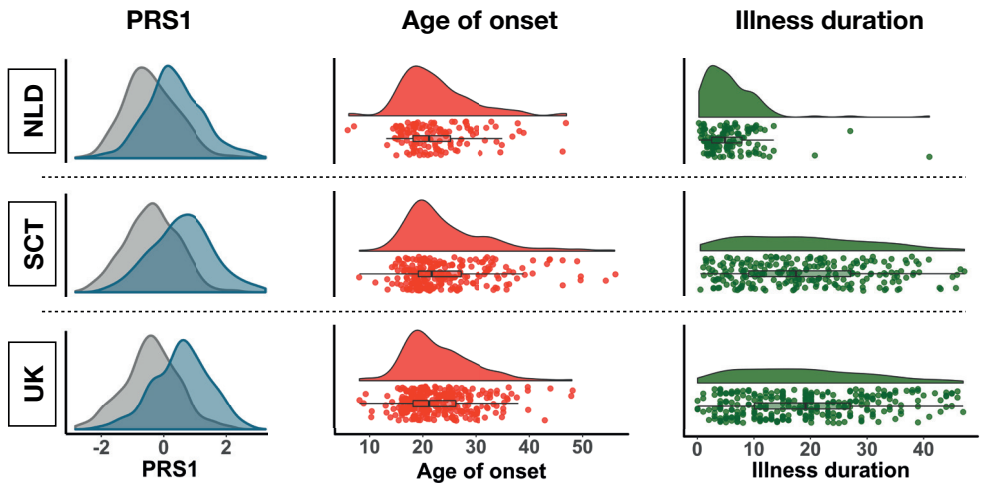


Figure S12. Variable distribution of PRS, age of presentation, and illness duration. *The distribution of SCZ PRS1 score (left; cases in blue), age of presentation (middle), and illness duration (right) for each of the three cohorts with available information.*

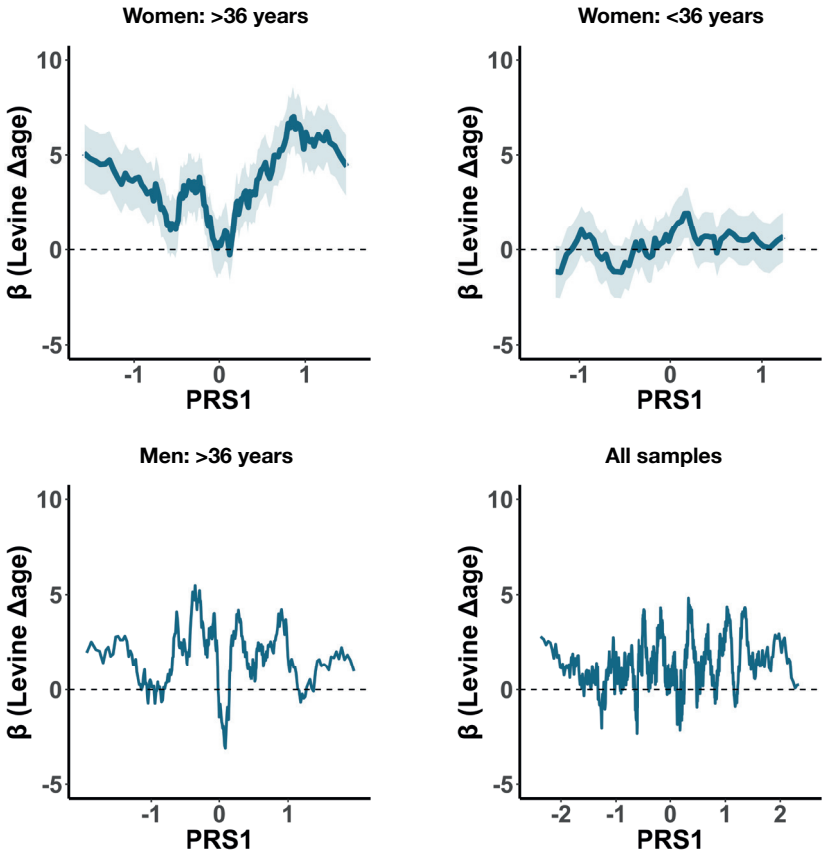


Figure S13. The association between DNAm aging and PRS1 is unique to women > 36. Using a sliding-window approach, Δage difference between cases and controls are shown across bins of ranked PRS1. Each bin contains 20 cases and slides from low to high PRS1 per shifts of one sample. The Levine Δage effect in each bin is shown in blue with the standard error in shaded blue. If the standard error is not shown, it was dropped to increase visual clarity. Results are presented for women > 36 (top left), women < 36 (top right), men > 36 (bottom left), and all samples (bottom right).

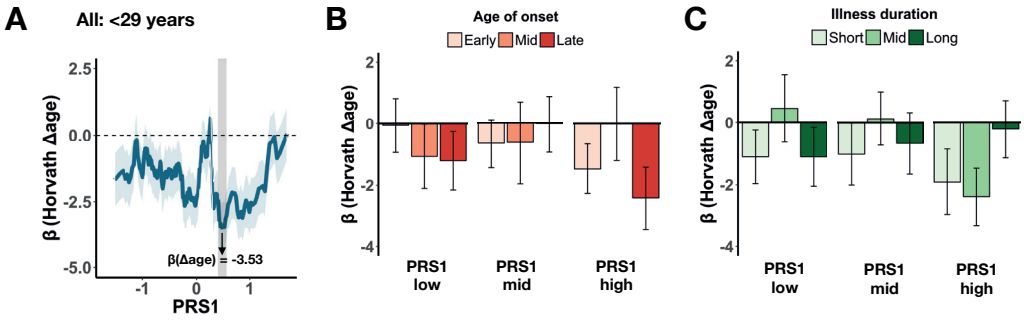


Figure S14. Integration of DNAm aging with PRS, age of presentation, and illness duration across identified age intervals. *(A) Using a sliding-window approach, Horvath Δage difference between cases and controls are shown across bins of ranked PRS1. Each bin contains 20 cases and slides from low to high PRS1 per shifts of one sample. The estimated Δage difference compared to all male controls < 29 years is shown for each sliding bin in blue with the standard error in shaded blue. The most significant bin is highlighted by the grey vertical bar. (B) DNAm aging effects stratified by PRS1 and age of onset. (C) DNAm aging effects stratified by PRS1 and illness duration.*

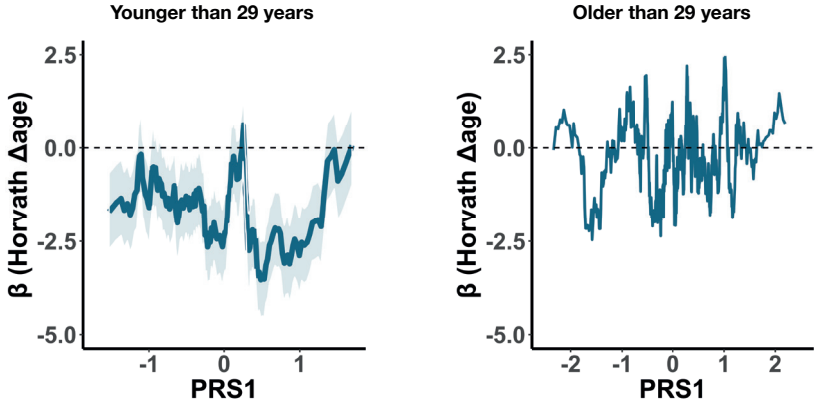


Figure S15. The association between Horvath Δage and PRS1 is more pronounced <29 years. *Using a sliding-window approach, Δage difference between cases and controls are shown across bins of ranked PRS1. Each bin contains 20 cases and slides from low to high PRS1 per shifts of one sample. The Horvath Δage effect in each bin is shown in blue with the standard error in shaded blue. If the standard error is not shown, it was dropped to increase visual clarity. Results are presented for all samples < 29 (left) all samples > 29 years (right).*

Supplementary Tables

	NLD	SCT	SWD	UK	Total
Samples (N)	926	665	69	636	2.296
Cases	461	260	37	332	1.090
Controls	465	405	32	304	1.206
Age (years)	35.8 (13.2)	44.6 (12.9)	59.7 (5.94)	40.4 (15.0)	40.3 (14.4)
Cases	35.5 (13.3)	44.2 (14.1)	59.6 (6.30)	43.7 (14.6)	40.9 (14.8)
Controls	36.1 (13.2)	44.9 (12.2)	59.9 (5.58)	36.8 (14.7)	39.8 (14.1)
Females	309 (33.4%)	185 (27.8%)	39 (56.5%)	259 (40.7%)	792 (34.5%)
Cases	122 (26.5%)	83 (31.9%)	20 (54.1%)	90 (27.1%)	315 (28.9%)
Controls	187 (40.2%)	102 (25.2%)	19 (59.4%)	169 (55.6%)	477 (39.6%)
Hannum DNAm age	37.9 (15.4)	46.0 (14.1)	61.7 (8.37)	41.6 (15.7)	43.1 (15.6)
Hannum Δage	0.0 (4.7)	1.4 (4.2)	2.0 (4.4)	1.2 (4.7)	1.0 (4.6)
Horvath DNAm age	37.0 (14.0)	47.6 (13.5)	62.0 (7.7)	41.2 (14.5)	42.0 (15.0)
Horvath Δage	1.2 (4.8)	3.0 (4.7)	3.2 (4.5)	0.9 (5.7)	1.7 (5.12)
Levine DNAm age	27.5 (15.1)	37.6 (15.4)	52.9 (8.1)	32.6 (16.5)	32.6 (16.4)
Levine Δage	-8.33 (6.7)	-7.0 (6.2)	-6.8 (5.4)	-7.8 (6.6)	-7.7 (6.5)

Table S1. Sample characteristics and DNA methylation age estimates across cohorts. Sample characteristic and mean values of chronological age and DNAm age estimates of each clock are presented for each cohort after quality control. Δage is defined by subtracting chronological age from DNAm age. Standard deviations are in parentheses unless otherwise defined. NLD = the Netherlands, SCT = Scotland, SWD = Sweden, UK = United Kingdom.

Dataset	PI/Contact	Ancestry	Platform	Data type used	Total	Cases	Controls	Age (sd)
GSE4103	RA Ophoff	Dutch	27K	IDAT files	624	337	287	33.3 (12.1)
GSE4119	RA Ophoff	Dutch	450K	IDAT files	96	62	34	31.1 (10.2)
TBD	RA Ophoff	Dutch	450K	IDAT files	324	160	164	34.4 (11.4)
TBD	RA Ophoff	Dutch	450K	IDAT files	72	36	36	59.2 (5.7)
TBD	PF Sullivan	Swedish	450K	IDAT files	96	48	48	59.8 (5.9)
GSE80417	A McQuillin	Scottish	450K	(un)methylated intensities	847	414	433	44.6 (12.9)
GSE84727	D St. Clair	UK	450K	(un)methylated intensities	675	353	322	40.4 (15.0)
Total		EUR	27/450K	Mixed	2.734	1.410	1.324	40.4 (28.1)

Table S2. Overview of datasets included in study. Multiple datasets of whole blood DNAm data across four European cohorts were included in the study. Shown above are some sample characteristics and accompanying GEO accession numbers for each dataset before quality control. PI = Principal Investigator, UK = United Kingdom, EUR = European.

Dataset	Ancestry	Platform	Tissue	Data type	Total	Cases	Controls	Age (sd)
GSE74193	AA/EUR	450K	DLPFC	IDAT files	503	224	279	46.9 (15.4)
GSE61107	-	450K	Frontal cortex	IDAT files	48	24	24	61.7 (19.2)
GSE61380	-	450K	Frontal cortex	(un)methylated intensities	33	18	15	44.0 (15.7)
GSE61431	-	450K	Frontal cortex	(un)methylated intensities	43	20	23	61.8 (17.5)
Total	Mixed / unknown	450K	Frontal cortex	Mixed	627	286	341	

Table S3. Overview of datasets included in brain analysis. Multiple datasets of postmortem brain DNAm data were included in in the analysis. Shown above are some sample characteristics and accompanying GEO accession numbers for each dataset before quality control. AA = African American, EUR = European, DLPFC = Dorsolateral prefrontal cortex.

lm(formula = Δage ~ Dataset + Ethnicity + Platform + Age.continuous + Status*Age.groups*Sex)

	Horvath Δage				Levine Δage			
Model variables	**Df**	**Sum Sq**	**Mean Sq**	**P-value**	**Df**	**Sum Sq**	**Mean Sq**	**P-value**
Dataset	6	1795	299,2	2,0E-15	6	679	113,2	6,7E-03
Ethnicity	-	-	-	-	-	-	-	-
Platform	-	-	-	-	-	-	-	-
Age.continuous	1	116	115,8	2,1E-02	1	1054	27,7	1,5E-07
Sex	1	46	45,7	1,5E-01	1	412	412,2	1,0E-03
Age.group	4	645	161,3	6,1E-06	4	512	128,0	9,3E-03
Status	1	288	288,0	2,8E-04	1	916	915,7	9,8E-07
Age.group:Status	4	227	56,7	3,4E-02	4	491	122,7	1,2E-02
Status:Sex	1	12	11,9	4,6E-01	1	51	50,9	2,5E-01
Age.group:Sex	4	329	82,1	4,5E-03	4	694	173,5	1,1E-03
Age.group:Sex:Status	4	66	16,3	5,5E-01	4	273	68,3	1,3E-01
Residuals	2135	46331	21,7	-	2137	81187	38,0	-

Table S9. Results three-way interaction model of age, sex, and status on Δage. *Shown are the contributions of each variable in the three-way interaction model presented by an analysis of variance table. The full model is displayed in the top row. Age.groups are defined by decades. Ethnicity and Platform are collinear with Dataset and thus do not have output. Df = degrees of freedom; Sum Sq; sum of squares; Mean Sq; mean of squares; P-value corresponds to the F-test in the anova() function.*

Sex	Age interval	Controls	Cases	Clock	Direction	β (Δage)	95% CI	P
Women	**<25**	84	28	Horvath	decelerated	-2,36	-4.07 — -0.64	7,3E-03
Women	**25-29**	75	28	Levine	accelerated	3,12	0.67 — 5.64	1,3E-02
Women	**>36**	217	192	Levine	accelerated	3,21	1.93 — 4.50	1,3E-06
Women	**43-47**	41	29	Horvath	decelerated	-3,05	-5.07 — -1.04	3,5E-03
Women	**50-56**	45	32	Horvath	accelerated	2,96	0.69 — 5.24	1,1E-02
Men	**27-39**	206	249	Levine	accelerated	1,72	0.62 — 2.81	2,3E-03
Men	**<29**	187	223	Horvath	decelerated	-1,39	-1.11 — 1.37	3,6E-03
Men	**44-48**	62	45	Horvath	decelerated	-2,10	-4.03 — -0.17	3,3E-02
Men	**49-59**	126	110	Levine	accelerated	2,44	0.67 — 4.21	7,1E-03

Table S10. Sex-specific DNAm aging in schizophrenia is variable across chronological age. For each identified change point and corresponding age interval, the number of samples and estimated Δage difference between cases and controls (β) with corresponding 95% confidence intervals (CI), and p-value (P) are presented. The sex, type of clock, and direction of aging effect are shown as well.

		Horvath Δage		Levine Δage	
Model variables	**Model comparison**	**Δage R^2**	**P-value**	**Δage R^2**	**P-value**
Model x: baseline		3,9%	-	3,1%	-
Model y: baseline (+ smoking)	-	4,9%	-	5,3%	-
Model z: baseline (+ cell types)	-	8,2%	-	22,1%	-
Model 0: baseline (+ smoking/cell)	-	9,4%	-	22,8%	-
Model 1: + status	Model 0 vs 1	9,3%	0,86	22,8%	0,26
Model 2: + status*age.continuous	Model 1 vs 2	9,4%	0,10	23,2%	1,3E-03
Model 3: + status*age.groups	Model 1 vs 3	10,6%	2,1E-04	23,2%	0,05
Model 4: + status*age.groups*sex	Model 3 vs 4	10,8%	0,15	23,6%	0,05

Table S11. Age- and sex-specific effects of DNAm aging in schizophrenia adjusted for smoking and cell type estimastes. Shown are the contributions of interaction effects between disease status and age and sex on Δage when adjusted for DNAm smoking scores (baseline model y) and blood cell type proportions (baseline model z). The full baseline model is defined as Δage ~ dataset + ethnicity + age.continuous + sex + DNAm smoking score + DNAm blood cell type proportions. For other models, the variable(s) in addition to the full baseline variables are shown with the corresponding variance explained (R.) in Δage. Interaction terms with chronological age are modeled as a continuous variable (age.continuous) or a categorical variable (age.groups). The latter uses previously defined decades. Model comparison is performed to assess if the contribution of an interaction term is significant compared to a model without that term. The chi-square test is used to test two models with corresponding p-value presented. The results of these analysis are shown for both the Horvath and Levine clock. These analyses included only 450K samples for which smoking scores and cell type estimates can be computed (N=1,621, 867 controls and 754 cases).

Women: >36 years (case = 165, control = 178)	**Variable R2**	**Variable P**	**Variable R2 adjusted**	**P adjusted**
Model - all selected variables	23,6%	2,2E-08	-	-
Levine Δage	**7,7%**	**6,0E-06**	**2,8%**	**3,3E-03**
Batch/Ethnicity	1,6%	5,2E-01	2,9%	1,1E-01
Smoking	9,3%	6,7E-07	4,7%	1,6E-04
CD8.naive	0,1%	5,8E-01	1,0%	7,4E-02
CD4.naive	2,6%	9,5E-03	0,2%	4,6E-01
CD8T	6,2%	5,5E-05	0,3%	3,7E-01
NK	6,2%	5,1E-05	0,0%	7,2E-01
Granulocytes	8,7%	1,5E-06	0,7%	1,3E-01

All samples: <29 years (case = 152, control = 146)	**Variable R2**	**Variable P**	**Variable R2 adjusted**	**P adjusted**
Model - all selected variables	49,8%	3,7E-32	-	-
Horvath Δage	**3,1%**	**2,1E-03**	**0,6%**	**1,4E-01**
Batch/Ethnicity	7,4%	3,6E-05	0,6%	5,4E-01
Age	0,0%	8,3E-01	0,2%	3,9E-01
Sex	8,9%	1,2E-07	1,0%	2,7E-02
Smoking	28,8%	3,3E-23	18,6%	2,3E-19
CD8T	3,7%	7,8E-04	0,0%	9,7E-01
Granulocytes	3,4%	1,3E-03	0,4%	1,7E-01
CD8pCD28nCD45RAn	5,3%	5,3E-05	1,0%	3,1E-02
PlasmaBlast	0,2%	4,3E-01	0,6%	7,8E-02
NK	7,8%	7,9E-07	0,4%	1,5E-01
Bcell	3,1%	1,9E-03	0,0%	7,3E-01
Mono	2,8%	3,2E-03	0,5%	1,1E-01

Table S12. DNAm aging significantly contributes to schizophrenia independent of smoking and cell types. *Shown are variables that significantly explain variance in SCZ disease status, selected by a penalized logistic regression analysis. The top and bottom table present results for women >36 years and all samples <29 years, respectively. Only samples assayed on the 450K platform were included as DNAm-based smoking scores and cell type proportions could be computed and included in the analysis. The top row of each table shows the proportion of variance explained in disease status (R^2) for all selected variables combined and the significance of a logistic regression model (glm, family="binomial") with each variable included compared to the null model of no variance explained. We also show the proportion of variance explained by each variable individually (Variable R2) and by each variable adjusted for all other selected variables (Variable R2 adjusted). The significance of Variable R2 adjusted is computed by comparing the model with all variables to a model with the variable of interest removed using the anova(test = "LRT") function. The result of this test is shown in the "P adjusted" column.*

All samples: <29	Controls	Cases	Mean value in cases	β (Horvath Δage)	95% CI	P
Polygenic risk						
All - no stratification	379	243	0,22	-1,30	-2.00 — -0.60	3,1E-04
PRS1 - continuous	-	243	0,22	-0,35	-0.77 — 0.07	1E-01
PRS1 - low	379	81	-0,77	-1,01	-2.05 — 0.02	5,6E-02
PRS1 - mid	379	81	0,17	-1,32	-2.36 — -0.27	1,7E-02
PRS1 - high	379	81	1,25	-1,58	-2.62 — -0.54	3,0E-03
Age of onset						
All - no stratification	379	190	19,48	-1,47	-2.23 — -0.71	1,6E-04
AOO - continuous	-	190	19,48	-0,03	-0.21 — 0.15	7,3E-01
AOO - early	379	64	15,91	-1,49	-2.61 — -0.37	9,8E-03
AOO - mid	379	63	19,19	-1,06	-2.23 — 0.10	7,2E-02
AOO - late	379	63	23,41	-1,85	-3.03 — -0.68	2,0E-03
Illness duration						
All - no stratification	379	190	4,92	-1,47	-2.23 — -0.71	1,6E-04
DUR - continuous	-	190	4,92	0,06	-0.14 — 0.25	5,7E-01
DUR - short	379	64	1,50	-1,86	-3.04 — -0.70	1,8E-03
DUR - mid	379	63	4,68	-1,31	-2.45 — -0.17	2,5E-02
DUR - long	379	63	8,63	-1,23	-2.38 — -0.08	3,6E-02

Table S13. Integration of Horvath Δage with PRS, age of onset, and illness duration in early adulthood. Analyses were performed by stratifying the analyses to only men and women <29 years of age. Only cases with available information were included in the analyses. Each phenotype was analyzed as both a continuous variable and as a categorical variable using equal tertiles from low to high bins. Mean values in cases for each phenotype are presented along with the association with Δage (β) and corresponding 95% confidence intervals and p-values. PRS1 = polygenic risk score PC1 (see Supplementary Information) scaled to mean zero with standard deviation of 1, AOO = age of onset, DUR = illness duration.

References – Supplementary Information

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." Bioinformatics 30 (10): 1363–69.

Bell, Jordana T., Pei-Chien Tsai, Tsun-Po Yang, Ruth Pidsley, James Nisbet, Daniel Glass, Massimo Mangino, et al. 2012. "Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population." PLoS Genetics 8 (4): e1002629.

Bergen, Sarah E., Alexander Ploner, Daniel Howrigan, CNV Analysis Group and the Schizophrenia Working Group of the Psychiatric Genomics Consortium, Michael C. O'Donovan, Jordan W. Smoller, Patrick F. Sullivan, Jonathan Sebat, Benjamin Neale, and Kenneth S. Kendler. 2019. "Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia." The American Journal of Psychiatry 176 (1): 29–35.

Christensen, Brock C., E. Andres Houseman, Carmen J. Marsit, Shichun Zheng, Margaret R. Wrensch, Joseph L. Wiemels, Heather H. Nelson, et al. 2009. "Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context." PLoS Genetics 5 (8): e1000602.

Datta, S. R., A. McQuillin, M. Rizig, E. Blaveri, S. Thirumalai, G. Kalsi, J. Lawrence, et al. 2010. "A Threonine to Isoleucine Missense Mutation in the Pericentriolar Material 1 Gene Is Strongly Associated with Schizophrenia." Molecular Psychiatry 15 (6): 615–28.

Eijk, Kristel R. van, Simone de Jong, Eric Strengman, Jacobine E. Buizer-Voskamp, René S. Kahn, Marco P. Boks, Steve Horvath, and Roel A. Ophoff. 2015. "Identification of Schizophrenia-Associated Loci by Combining DNA Methylation and Gene Expression Data from Whole Blood." European Journal of Human Genetics: EJHG 23 (8): 1106–10.

Elliott, Hannah R., Therese Tillin, Wendy L. McArdle, Karen Ho, Aparna Duggirala, Tim M. Frayling, George Davey Smith, Alun D. Hughes, Nish Chaturvedi, and Caroline L. Relton. 2014. "Differences in Smoking Associated DNA Methylation Patterns in South Asians and Europeans." Clinical Epigenetics 6 (1): 4.

Hannon, Eilis, Emma Dempster, Joana Viana, Joe Burrage, Adam R. Smith, Ruby Macdonald, David St Clair, et al. 2016. "An Integrated Genetic-Epigenetic Analysis of Schizophrenia: Evidence for Co-Localization of Genetic Associations and Differential DNA Methylation." Genome Biology 17 (1): 176.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." Molecular Cell 49 (2): 359–67.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." Genome Biology 14 (10): R115.

Horvath, Steve, Yafeng Zhang, Peter Langfelder, René S. Kahn, Marco P. M. Boks, Kristel van Eijk, Leonard H. van den Berg, and Roel A. Ophoff. 2012. "Aging Effects on DNA Methylation Modules in Human Brain and Blood Tissue." Genome Biology 13 (10): R97.

Houseman, Eugene Andres, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. 2012. "DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution." BMC Bioinformatics 13 (1): 86.

International Schizophrenia Consortium. 2008. "Rare Chromosomal Deletions and Duplications Increase Risk of Schizophrenia." Nature 455 (7210): 237–41.

Jaffe, Andrew E., and Rafael A. Irizarry. 2014. "Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies." Genome Biology 15 (2): R31.

Kelly, Ciara, and Robin McCreadie. 2000. "Cigarette Smoking and Schizophrenia." Advances in Psychiatric Treatment 6 (5): 327–31.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." Aging 10 (4): 573–91.

Niu, Liang, Zongli Xu, and Jack A. Taylor. 2016. "RCP: A Novel Probe Design Bias Correction Method for Illumina Methylation BeadChip." Bioinformatics 32 (17): 2659–63.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." Nature 511 (7510): 421–27.

Triche, Timothy J., Jr, Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. 2013. "Low-Level Processing of Illumina Infinium DNA Methylation BeadArrays." Nucleic Acids Research 41 (7): e90.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses inRwith themetaforPackage." Journal of Statistical Software 36 (3).

Xu, Zongli, Sabine A. S. Langie, Patrick De Boever, Jack A. Taylor, and Liang Niu. 2017. "RELIC: A Novel Dye-Bias Correction Method for Illumina Methylation BeadChip." BMC Genomics 18 (1): 4.

Xu, Zongli, Liang Niu, Leping Li, and Jack A. Taylor. 2015. "ENmix: A Novel Background Correction Method for Illumina HumanMethylation450 BeadChip." Nucleic Acids Research 44 (3).

Zeilinger, Sonja, Brigitte Kühnel, Norman Klopp, Hansjörg Baurecht, Anja Kleinschmidt, Christian Gieger, Stephan Weidinger, et al. 2013. "Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation." PloS One 8 (5): e63812.

PART 3

Discussion of results and conclusions

CHAPTER 7

Discussion of research findings and conclusions

We are in a promising era of human genomic and psychiatric research. Not only did we identify regions in the genome and, in some instances, even the underlying genes that confer risk for psychiatric disorders for the first time, but we are also rethinking the phenotypic domains and the classification systems in which clinical diagnoses of psychiatric disorders exist. Our understanding of the molecular causes and consequences of psychiatric disorders is currently, however, still limited and no cures are available yet. Clarifying the underlying biological mechanisms of psychiatric disorders is a necessary step towards the development of new treatments and the further humanization of these devastating and stigmatized illnesses. Genomics has proven to be an instrumental approach for studying the biology of complex human traits, including schizophrenia, a major psychiatric disorder that affects millions of people worldwide. Large-scale genetic studies have, for example, been successful at identifying genetic variations associated with schizophrenia, thereby propelling new insights into its pathogenic mechanisms. The next step forward is to conduct more mechanistic studies that build upon these genetic associations to clarify the molecular and cellular processes that are disrupted. On the other hand, understanding how non-genetic determinants and molecular consequences of the illness contribute to its pathophysiology is equally important.

My dissertation had the broad aims of: (1) translating the findings from GWAS into disease biology by investigating the functional mechanisms that underlie schizophrenia heritability, (2) mapping the molecular profile of clozapine response by genome-wide gene expression and DNA methylation data analyses, and (3) investigating the molecular consequences of the illness by quantifying biological age using DNAm clocks. These aims are embedded in two sections of the dissertation. The first part focuses on understanding the molecular biology of schizophrenia by using *in vitro* experimental systems, while the second part of the dissertation investigates the reliability DNAm-based predictors and the role of DNAm age in schizophrenia using DNA methylation clocks. My research primarily focuses on schizophrenia, but its findings and implications have broad relevance for other psychiatric illnesses and complex genetic traits in general. Next, I discuss the findings and conclusions of my research by summarizing them in eight take-home messages that I believe to be important. Finally, I discuss the limitations of the work and its implications for future research.

Part 1: Functional investigations of schizophrenia biology

Take-home 1: embrace polygenicity in functional investigations of schizophrenia

Schizophrenia is a complex genetic trait with a highly polygenic disease architecture. How best to integrate aggregate disease risk – i.e., the genetic susceptibility that is distributed across hundreds to thousands of alleles – with *in vitro* experimental systems is a largely open question in the post-GWAS era, particularly for psychiatric disorders. Functional experiments of single genes or genetic variants can be powerful in demonstrating molecular consequences, i.e., assuming the causal variant has been identified, but fail to take into account the polygenic nature of the illness. Furthermore, gene silencing or full gene knockouts may induce a downstream effect in gene function that could be substantially off compared to the functional consequences induced by the disease-associated allele. Incorporating polygenicity into the experimental design of follow-

up functional investigations of genetic risk of complex human traits may, therefore, offer insights beyond what studies of single genes or genetic variants in isolation of the rest of the genome can yield.

I conducted a functional investigation of schizophrenia genetic susceptibility by using an *in vitro* stem cell model of human neuronal differentiation, detailed in Chapter 3. This *in vitro* experimental model captures specific spatiotemporal dimensions of early stages of development in the human brain. Using an analytical approach tailored to time-series data, I identified eight longitudinal gene clusters that are involved in the development of neurons. An important aspect of this work was to integrate these gene clusters with the heritability of schizophrenia that is measured by the GWAS. Using statistical methods that are tailored to modeling genome-wide disease risk, I integrated schizophrenia polygenic risk with longitudinal transcriptomic signatures of human neuronal differentiation. What I found is that schizophrenia polygenic risk is significantly enriched in genes that are differentially expressed during neuronal differentiation, and more specifically, that this enrichment is driven by a specific longitudinal gene cluster of genes involved in synaptic functioning. To the best of my knowledge, this is the first demonstration of how parts of the heritability of schizophrenia can be detected in an *in vitro* experimental model of human neuronal differentiation.

Finding appropriate model systems for functional investigations of psychiatric disorders is a significant challenge. This finding is therefore of value, as it provides validity for using *in vitro* human neuronal differentiation as a model to study genetic susceptibility of schizophrenia. It is important to emphasize that the genes that contribute to the observed enrichment of schizophrenia genetic risk likely only capture parts of the measured heritability of the GWAS. Genes involved in other cell types or different developmental stages, and that are therefore not measured or variable in this experiment, could carry genetic risk as well. Schizophrenia heritability has, for example, been shown to map on several brain cell types, including pyramidal cells, medium spiny neurons, and interneurons (Skene et al. 2018). While the cellular identity of the *in vitro* neuronal differentiation model I used was broadly neuronal, I did not establish more in-depth classification by cellular taxonomy of the brain. It could be true that a different differentiation protocol would change the outcome of the heritability analysis. Tweaking the experimental parameters of the model will help further understand how downstream molecular changes are associated with schizophrenia genetic risk and help in the search for what genes and pathways underlie the observed heritability. In the figure below, I present a conceptual framework of what to think about when using GWAS heritability to map disease biology in follow-up functional investigations.

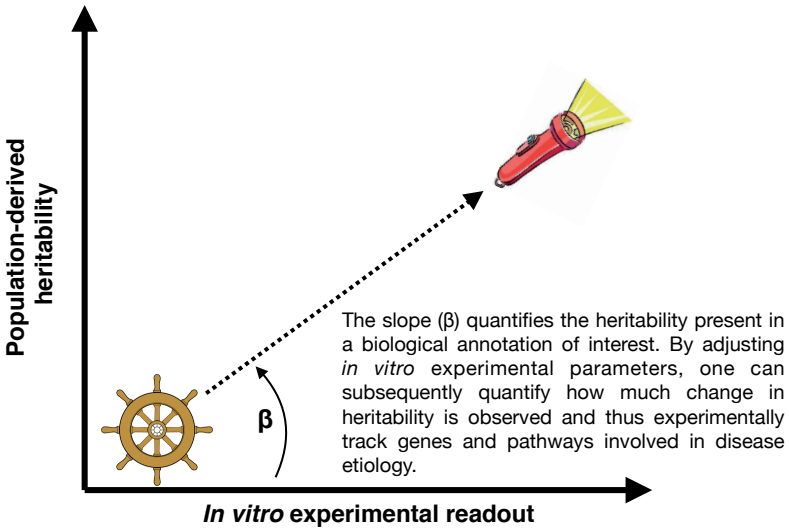


Figure 1. Modeling heritability in a dish. *Shown is a conceptual presentation of a bioinformatic framework that integrates GWAS heritability with functional annotations obtained through experimental lab models. Both the x- and y-axis can be updated, for example when GWAS sample sizes increase (y-axis) or when new experiments are conducted in the lab (x-axis). The model can be used to scan the available functional space, to the extent the model can capture, and "track" disease heritability to study the biology that underlies the trait of interest in an unbiased fashion.*

In my work, I choose a 30-day *in vitro* human neuronal differentiation of an isogenic neural stem cell line as an experimental model and a genome-wide gene expression as a functional readout. An important question is how the enrichment of schizophrenia heritability changes when one matures the culture for a longer period or differentiates to a more specific neuronal lineage or a different brain cell type, like astrocytes or oligodendrocytes. In addition to gene expression, one can also use epigenetic, proteomic, or other molecular and cellular units as readouts. The more dimensions and functional layers we analyze, the more light will be shed on (1) our understanding of the molecular biology that underlies aggregate genetic risk of schizophrenia, and (2) how specific the observed enrichment of synaptic genes is for the model of human neuronal differentiation. I see such an analysis framework as a stepping-stone to move from genetic susceptibility, identified in the GWAS, to a single-nucleotide resolution understanding of schizophrenia biology across the genome. Once a specific spatiotemporal dimension and experimental model is identified to be relevant for studying the genetic susceptibility of a trait of interest, more high-throughput, labor heavy, and/or cost-heavy assays can be performed, such as single cell analyses or massively parallel reporter assays (MPRAs). MPRAs, for example, functionally screen thousands of genetic sequences for regulatory activity in parallel and have been successfully applied in understanding enhancer activity (Klein et al. 2020; Patwardhan et al. 2012). Future work could perform a MPRA of schizophrenia-associated genetic sequences in a model of human neuronal differentiation to understand how these individual genetic variants

contribute to the observed heritability enrichment in the synaptic gene cluster. Similar analysis of neuronal differentiation using induced pluripotent stem cells of individuals with a schizophrenia diagnosis or a high burden of schizophrenia genetic risk could be an important avenue to explore as well (Buxbaum et al. 2019). In a study that I co-authored, we identified that schizophrenia cases carry an increased burden of low frequency deleterious mutations across the genome compared to controls (Olde Loohuis et al. 2015). Extending functional investigations to capture the collective burden of low frequency to rare variants and subsequently integrating this with genome-wide polygenic risk, should be explored as well in future research to further fine-map schizophrenia genetic susceptibility.

The opportunities to expand on this work really are abundant. The challenge is to carefully select and think through an experimental plan when there are many decisions to make and parameters to set. Heritability analysis of common genetic variation can serve as a complementary approach to already existing laboratory techniques and assays to search for and map disease biology. By leveraging the success of the GWAS and integrating heritability analysis into functional investigations, we also embrace the highly polygenic nature that is inherent to psychiatric disorders in our search for clarity.

Take-home 2: schizophrenia genetic risk maps to synaptic biology but higher functional resolution is needed

GWAS has uncovered hundreds of genetic variants that influence risk of schizophrenia with new discoveries continuing to be made as the sample sizes grow and new cohorts are included (Consortium et al., n.d.). While the mapping of causal variants to pathogenic mechanisms remains challenging for any complex human trait (Gallagher and Chen-Plotkin 2018), the evidence in schizophrenia so far does strongly suggest that synaptic biology and plasticity are implicated. Synapses are intercellular junctions specialized in fast, cell-to-cell information transfer in the brain (Südhof 2018). Synapses are at the heart of brain plasticity and shape the development of the human brain. Both small and large de novo mutations identified in individuals with schizophrenia diagnoses show enrichment in synaptic genes, in particular genes encoding proteins of the N-methyl-D-aspartate (NMDA) receptor (NMDAR complex) and proteins that interact with the activity-regulated cytoskeleton-associated protein (ARC complex) (Glessner et al. 2010; Malhotra et al. 2011; Kirov et al. 2012; Fromer et al. 2014). The NMDAR and ARC complexes represent interconnected biological mechanisms at the postsynaptic density (PSD), which is a dense network of proteins embedded in the postsynaptic membrane of neurons (Hall et al. 2015).

These findings are further solidified by analysis of ultra-rare protein-altering variants (URVs) in a large case-control exome sequencing cohort (Genovese et al. 2016). This study observed an excess of disruptive URVs in synaptic genes, including the NMDAR and ARC complexes and in genes that bind to the fragile X mental retardation protein (FMRP) and the RNA Binding Fox-1 and Fox-3 proteins (RBFOX1/3). FMRP, RBFOX1, and RBFOX3 are RNA binding proteins that have been observed at synapses and regulate synaptic messenger RNAs and downstream synaptic function (Fernández, Rajan, and Bagni 2013; Lee et al. 2016). The authors were concerned that the enrichment of disrupted URVs in synaptic genes could simply be a result of the experiments that generated these synaptic gene sets and annotations conducted in neurons, which have a high

expression of synapse-associated genes. The enrichment signal could therefore be confounded by synaptic genes, which are also broadly neuronal genes. Genes expressed in neurons, and the central nervous system in general, are known to carry a larger burden of schizophrenia genetic risk (Finucane et al. 2018). A similar train of thought holds for my finding of schizophrenia heritability enrichment in a longitudinal gene cluster of synaptic genes. As I differentiated neural stem cells to a neuronal lineage, genes that are then activated have predominantly neuronal specificity. However, I found that even when I conditioned my enrichment analysis on the expression of other upregulated gene clusters, which consist of neuronal genes, the observed association with the synaptic gene clusters remained significant. This observation is similar to what the study of disrupted URVs in schizophrenia found when they performed their analysis separately for non-synaptic and synaptic neuronal genes. That is, the excess burden in synaptic gene sets was only seen in the synaptic neuronal genes and not in other neuronal genes. While this is an important observation, gene sets of synaptic protein complexes and RNA binding proteins still consist of hundreds to possibly even thousands of genes. More functional specificity is required to further disentangle how synaptic biology contributes to the pathogenic mechanism of schizophrenia.

To generate a more comprehensive definition of genes and cellular processes at the synapse, the Synaptic Gene Ontologies (SynGO) database was established. SynGO is a recently established public data resource that contains evidence-based expert-curated annotations for synaptic function and processes (Koopmans et al. 2019). SynGO contains almost 3,000 annotations against 1,112 experimentally validated synaptic genes. While thousands of synaptic genes are still awaiting validation, SynGO currently represents the most carefully curated and detailed classification of genes involved in synaptic biology. SynGO analysis demonstrated that de novo mutations found in schizophrenia were also enriched in SynGO genes, confirming results described above. This is in line with SynGO genes being highly intolerant to loss-of-function mutations. The latest schizophrenia GWAS also performed heritability enrichment analysis on SynGo ontologies and confirmed strong common genetic variant associations with postsynaptic terms, but also found associations with presynaptic and transsynaptic genes and processes (Consortium et al., n.d.). Together, this demonstrates that studies of both rare and common genetic variant associations converge on the finding that schizophrenia pathophysiology concentrates at the synapse for a large part.

Synaptic genes are not confined to one cellular compartment or a single biological process but are intricately connected in their function. Further understanding of schizophrenia pathogenicity will require a finer functional resolution of synapse biology. My finding that schizophrenia polygenic risk is enriched in a synaptic gene cluster of an *in vitro* cell-based experimental model is therefore timely and important. Efforts to prioritize likely causal genes have yielded several synaptic genes that show converging evidence of genetic association across multiple study designs (Consortium et al., n.d.). These genes should be primary candidates for follow-up mechanistic studies. In-depth functional investigations of empirically supported genetic associations can yield valuable insights into disease pathogenicity. One such study, for example, found that genetic variation in the major histocompatibility complex locus, representing the strongest genetic association with schizophrenia at population level, in part stems from distinct alleles of the complement component 4 (C4) genes (Sekar et al. 2016). They

showed that the common C4 genetic variation associates with schizophrenia and increases expression in multiple brain regions and localizes to dendrites and synapses in neurons. In mice, C4 is involved in synaptic pruning during postnatal development. How C4 function relates *in vitro* human neuronal differentiation and to the activity of the synaptic gene cluster remains to be answered. Further mapping of the genetic susceptibility of schizophrenia on specific molecular and regulatory processes involved in synaptic biology will be pivotal.

Future work on schizophrenia pathogenicity that aims to use *in vitro* human neuronal differentiation as an experimental model would benefit from a robust quantification of synapse formation and strength. Activity-dependent synaptic activity is important for brain development and a defining property of neurons (Chaudhury et al. 2016; Rebola, Srikumar, and Mulle 2010) I did not perform any quantification of cellular morphology or neurophysiological properties in my investigations. As experimental conditions, like culture media, are known to affect the neurophysiological properties of neurons (Bardy et al. 2015), more work is needed to assess how this impacts the observed findings. *in vitro* co-cultures of neurons and astrocytes could be an avenue to explore. Astrocyte co-cultures exert pro-maturational effects on neuronal cells, especially on the development of synapses (Hedegaard et al. 2020; Banker 1980). Astrocytes have important roles in regulating neurotransmitter and general synapse homeostasis and promote neurite outgrowth for example. Performing neuronal differentiation with co-cultured astrocytes could therefore yield greater activity of synapses and more diversity in genes as well.

Finally, alterations in synaptic biology have been implicated in other psychiatric disorders as well (Koopmans et al. 2019). As genetic influences on psychiatric disorders transcend diagnostic boundaries (Cross-Disorder Group of the Psychiatric Genomics Consortium. 2019), dissecting shared and disorder-specific pathogenic mechanisms will be pivotal for clarifying the biology of schizophrenia. The recent developments of genomic structural equation modeling allow for joint analysis of genetic architecture of complex traits (Grotzinger et al. 2019), and could be an avenue for exploring and identifying schizophrenia-specific polygenic risk. For now, schizophrenia genetic susceptibility maps strongly onto synaptic biology, but it may be that this is the rule and not an exception in general psychiatric pathophysiology.

Take-home 3: Lymphoblast cell lines make for a useful experimental tool to study adverse molecular medication effects associated with clozapine

Unraveling the pathogenic mechanisms of schizophrenia can help us understand the molecular chain of events that cause the illness and thereby accelerate the development of new therapeutics. While pathogenic mechanisms are central to the biology of the illness, they are not the only molecular and cellular events that impact the health of patients. Individuals with a diagnosis of schizophrenia, for example, also suffer from the biological consequences of adverse antipsychotic medication effects (Gonçalves, Araújo, and Martel 2015). These adverse effects can drastically reduce quality of life, even with life-threatening consequences in some cases, and significantly lower adherence to pharmacological treatment as well (Dibonaventura et al. 2012). To improve our understanding of antipsychotic medication, and thereby our ability to provide the necessary medical care for patients, more clarity on the molecular consequences of these drugs is needed.

In Chapter 4, I investigate the molecular signatures of clozapine response by implementing a human lymphoblast cell line model in which cells were exposed to increasing doses of the antipsychotic drug. While clozapine is one of the most effective antipsychotic drugs in reducing symptoms of schizophrenia and has good drug adherence among antipsychotics, it can induce strong secondary outcomes (Leucht et al. 2013). In rare instances, the use of clozapine may lead to agranulocytosis, for example (Andersohn, Konzen, and Garbe 2007). Far more frequent, however, are unwanted outcomes of weight gain and sedation (Leucht et al. 2013; Baptista 1999). These adverse effects are likely to contribute to serious long-term cardiometabolic consequences, such as diabetes and cardiovascular diseases, which are reported at higher rates in patients diagnosed with schizophrenia (Galletly et al. 2012; Olfson et al. 2015). So far, the molecular mechanisms underlying clozapine response are not well understood. Using an *in vitro* lymphoblast cell line model, I therefore examined genome-wide gene expression and DNAm changes in response to clozapine exposure. LCLs were exposed to clozapine concentrations from 1x up to 100x clinical concentration, which is an extension of previous experimental lab work to investigate LCL cell viability in response to clozapine (de With et al. 2015). The rationale is that a wide range of concentrations will help to identify genes and DNAm sites that have a dose-response change in expression level. Using this experimental study design, the aim was to gain insights into the basic molecular mechanisms that underlie adverse effects of the drug by identifying genes and pathways that change in activity in response to drug exposure *in vitro*. Two novelties of the study are the genome-wide analysis across multiple genomic data layers and the state-of-the-art integrative analyses with other genomic resources, such as SNP-based heritability from GWAS and the GTEx human tissue gene expression dataset.

The results showed strong activation of hundreds of genes after exposure to clozapine. Genes upregulated by clozapine response are linked to cholesterol metabolism and steroid biosynthesis, while downregulated genes are involved in various cell division processes. The central role of cholesterol metabolism in the response to clozapine is concordant with previously reported candidate gene studies performed in other cell types and for other antipsychotic drugs as well (Fernø et al. 2005; Foley and Mackinnon 2014). My analysis of genome-wide gene expression changes after clozapine exposure demonstrates that transcriptomic consequences are widespread and are not limited to a handful of genes. As abundant as the changes in gene expression are, so few are the changes in levels of DNA methylation. CpG sites upstream of the low-density lipoprotein receptor (LDL-R) gene and the cyclin F (CCNF) gene did show a significant change in DNA methylation after 24 and 96 hours, respectively. These genes are involved in similar biological processes as the annotations related to changes in gene expression. The LDL-R gene is responsible for the uptake of cholesterol-carrying particles into the cell and is a central player in cholesterol metabolism (Fass et al. 1997), and the CCNF is involved in the coordination of essential cell cycle events (Bai, Richman, and Elledge 1994). How the changes in gene expression and DNA methylation are linked, if they are at all, is another avenue to explore.

To gain insights into how *in vitro* gene expression signatures of clozapine response translate to gene expression in human tissues, I performed both within and between tissue analysis in the GTEx gene expression dataset using the identified clozapine-associated genes as the starting point. The GTEx project is an ongoing effort to build a comprehensive public resource

for studying tissue-specific gene expression and regulation (Lonsdale et al. 2013). Across 22 human tissues, I found that clozapine-associated genes are most preferentially expressed in GTEx-LCL tissue. Whole blood tissue ranked as second best. This is not unexpected given that LCLs are blood-derived cells and show that features of the *in vitro* experiment translate to the GTEx dataset. Investigating all genes in GTEx, and not being limited to genes that are preferentially expressed in specific tissues, highlighted that clozapine genes have significantly different gene expression in multiple tissues, including the liver, muscle, lung, and testis. In addition to within tissue analysis, I also performed between tissue analysis and found tissue pairs related to immune (spleen), endocrine (testis, ovary, adrenal, and thyroid gland), and metabolic (adipose) functioning, which clozapine treatment may affect disproportionately. As endocrine and metabolic abnormalities are known causes of human obesity, the identified gene expression signature of cholesterol pathways in response to clozapine exposure together with downstream tissue effects may point to new avenues to study clozapine-induced weight gain using *in vitro* experimental studies. Other observed tissue effects, such as for endocrine organs like the testis and ovaries, warrants more research on the downstream effects of clozapine in these specific tissues.

Take-home 4: More questions than answers – there is still a lot we do not know about the molecular function of clozapine

My research on clozapine has yielded insights into the molecular response of the drug. While my findings point to specific genes, pathways, and tissues that could be implicated, these insights are merely the tip of the iceberg of how clozapine functions and raise important questions to address. Clozapine-associated genes were, for example, not differentially expressed in any brain tissue in GTEx, which is surprising given that clozapine is an antipsychotic drug that is used in treatment of brain disorders. The lack of association with brain tissues suggests that the LCL model may capture tissue-specific molecular signatures related to clozapine's adverse molecular effects, but not necessarily to genes and pathways related to its main therapeutic effects that yield symptom relief. That is, if clozapine's therapeutic effects are indeed mediated through brain cells. Gene expression analysis of human induced excitatory neurons at least supports upregulation of cholesterol biosynthesis pathways in response to clozapine as well (Das et al. 2021). How the main molecular effects of the drug differ from secondary unwanted outcomes is unclear. Performing similar *in vitro* experiments and genomic analysis in other cell types as well, including brain cell types, will shed more light on how generalizable the molecular signatures of clozapine response in LCLs are compared to other cell types. Multi-cell type analyses can then help differentiate molecular mechanisms of adverse and therapeutic effects in response to clozapine exposure.

To explore how clozapine-associated genes relate to the possible biology of schizophrenia, I assessed if differentially expressed clozapine genes overlap with genes that are associated with schizophrenia genetic risk. Previous studies have found an association between gene targets of antipsychotic drugs and the genetic susceptibility of schizophrenia (Gaspar and Breen 2017; Skene et al. 2018), which suggests that antipsychotic drug targets overlap with pathogenic mechanisms of the illness. If this is the case, the identified genes and pathways associated with clozapine may also give insights into causal mechanisms of the illness. To find an overlap between disease-associated

loci and targets of already-approved drug therapy would not be exclusive to schizophrenia and clozapine, as this has been reported for other illnesses as well. A well-known example is the GWAS of coronary artery disease that identified genetic loci that harbor gene targets of statins and other lipid-lowering medications (Tragante et al. 2018). However, I did not find such evidence for clozapine-associated genes and the schizophrenia GWAS. That is, genes up- or downregulated by exposure to clozapine do not carry more genetic risk of schizophrenia than genes that are not differentially expressed in lymphoblast cell lines. I did observe an enrichment of total cholesterol and LDL heritability in clozapine genes. While caution is warranted with the interpretation of this observation, it does align with the large number of cholesterol-associated genes that were found to be differentially expressed. Again, these findings suggest that clozapine-induced molecular profiles in lymphoblast cell lines may provide insights into metabolic adverse effects of the drug, but not necessarily mechanisms related to therapeutic effects or schizophrenia pathogenicity. It may well be that other studies observed an overlap between antipsychotic drug targets and schizophrenia genetic risk because their drug targets were identified in other tissue or cell types, by using other antipsychotic drugs or through in silico experiments, or because their findings are false positives. Overall, it remains an open question as to what the mechanistic functions of clozapine precisely are, both with regard to its effective actions and secondary outcomes. Given the high reported efficacy of the drug to reduce symptoms of schizophrenia, future work should investigate this further.

Clozapine and its metabolites have been associated with clozapine-induced agranulocytosis (Mijovic and MacCabe 2020; Bablenis, Weber, and Wagner 1989), which is an extreme form of neutropenia (lower-than-normal levels of white blood cells). Clozapine-treated people report a 0.4% incidence of agranulocytosis (Li et al. 2020). In high concentration, clozapine has been shown to affect the viability of LCLs (de With et al. 2015), which could be related to the downregulation of genes involved in cell cycle processes that are observed in this dissertation. In the between-tissue GTEx analyses, the spleen tissue significantly stood out as being impacted by clozapine-associated genes compared to other tissues. The spleen is the largest secondary lymphoid organ in the body and hosts a wide range of immunological functions, including the storage of leukocytes (Lewis, Williams, and Eisenbarth 2019). It is tempting to speculate that changes in cell cycle processes and splenic dysfunction upon clozapine exposure play a role in clozapine-induced agranulocytosis. This could serve as a plausible hypothesis to investigate in future research.

The findings of the genomic study on clozapine exposure should be interpreted in light of several limitations. First, clozapine concentrations ranging between 1x to 100x clinical concentration were used to induce a strong drug response and subsequent changes in downstream gene expression and DNAm levels. While the molecular profiles identified do capture key genes and biological pathways similar to those that have been reported by previous studies of clozapine response, it does remain an open question how these findings can be translated toward the clinic. Performing similar *in vitro* experiments but across a range of concentrations closer to the clinical concentration of clozapine and comparing those outcomes with the current findings is a next step. *In vivo* studies in human subjects are often costly and limited in their study designs. An important advantage of *in vitro* experimental models is the controlled laboratory

environment that allows for precise manipulation of the *in vitro* model. LCLs could serve as an experimental model system to study the molecular mechanisms of adverse effects further.

The results are likely biased by the *in vitro* cell type used and may include an overrepresentation of LCL-specific molecular effects. Studying one cell type in isolation will neither capture the complexity within human tissues nor the dynamics of cross-talk between tissues. This is a limitation of the study. Conducting similar *in vitro* analysis in other cell types or in three-dimensional tissue culture would be an important avenue to explore in future research on clozapine response. While my findings suggest that clozapine affects specific genes and pathways, and may disproportionately affect the function of specific tissues, they do not provide direct mechanistic insight yet. As the exact cascade of molecular and cellular events up clozapine exposure remains to be elucidated further, these findings can help formulate what next steps are needed to better understand the function of the drug. Future investigations should also take time-dependent effects of clozapine exposure into account, as my analysis showed that molecular profiles associated with clozapine exposure are different between 24 and 96 hours of drug exposure. This observation is similar to findings of a candidate gene expression study in human brain and liver cells that reported different patterns of expression of cholesterol-related genes depending on antipsychotic drug exposure time (Vik-Mo et al. 2009). Insights into time-dependent effects of clozapine are important because evidence of randomized controlled trials showed particularly strong efficacy for acute treatment of psychosis (Goff et al. 2017). How molecular changes in response to clozapine exposure compare to *in vitro* molecular profiles of other antipsychotic drugs is an open question as well.

Finally, it will be important to determine how the identified *in vitro* molecular changes upon clozapine exposure relate to molecular outcomes in patients treated with clozapine. A relatively small study that investigated whole blood tissue of 152 psychosis patients, of which 55 received clozapine, did not find significant changes in the expression of any genes (Harrison et al. 2016). This suggests that clozapine likely does not have large effects on gene expression in blood tissue and/or that a greater study sample size is needed to characterize its molecular consequences at the level of gene expression. There is evidence from studies of other psychiatric disorders that suggest that the inclusion of a larger number of individual may indeed be needed to study clozapine. In a study that I co-authored, we investigated gene expression in blood of people diagnosed with bipolar disorder using a larger case-control cohort of 480 individuals. We found that treatment with lithium medication had a strong effect on gene expression levels in the blood of patients and that changes due to the illness (and thus independent of medication use) were minimal (Krebs et al. 2020). While this study of lithium treated bipolar disorder patients does not provide evidence that larger cohorts of clozapine treated patients will identify significant molecular changes, it does suggest that pursuing such an avenue could be a worthwhile effort. *in vitro* functional analyses are by design limited in their translation to human biology. More human molecular studies are needed to study the function of clozapine in patients that are treated with the drug.

To summarize, my research on clozapine response identified specific genomic signatures that are associated with the antipsychotic drug and provide insights into possible adverse molecular consequences of clozapine. My findings were obtained using an *in vitro* experimental

model system and the results should be interpreted within the limitations of the study design and laboratory model system. While *in vitro* studies such as that employed in this dissertation may only capture elements of what may be relevant *in vivo*, they can serve as very useful tools, particularly when their output is integrated with external human genomic datasets, as demonstrated in my work. As the necessary challenge of large-scale, systematic prospective patient cohort studies of adverse effects of clozapine and other AP remains, genomic research in LCLs offers a complementary strategy towards understanding the molecular consequences of clozapine further.

Part 2: DNA methylation algorithms and the role of biological age in schizophrenia

Part 1 of my dissertation focused on functional investigations of schizophrenia in which I studied biological pathways that underlie the genetic etiology of the illness and possible molecular consequences of antipsychotic medication. In Part 2 of my dissertation, I investigate the role of DNAm age in schizophrenia in a large case-control cohort. DNAm is a type of epigenetic modification that is known to be influenced by both genetic and environmental factors (Hannon et al. 2018), and has been associated with schizophrenia (Mill et al. 2008; Hannon et al. 2016; Jaffe et al. 2015). New developments in epigenetic research have introduced DNAm-based algorithms as quantitative tools to study health and disease. DNAm clocks, which estimate biological age, are of particular interest as they have been associated with various diseases, including psychiatric illnesses and all-cause mortality (Horvath and Raj 2018). Early epidemiological reports have described schizophrenia as a life-shortening illness (Allebeck 1989), an observation that has now been confirmed by larger epidemiological studies. Individuals diagnosed with schizophrenia suffer disproportionately from age-related disabilities, report two-to-three times excess mortality, and have a lifespan of 15 years reduced by compared to the general population (McGrath et al. 2008; Olfson et al. 2015).

It remains an open question whether higher rates of age-related disabilities and morbidities reported in patient populations are a consequence of the illness or whether processes of innate aging are part of the etiology of schizophrenia. The molecular quantification of a changed aging process in schizophrenia could be an important step towards measuring the degree of risk for age-related disability and morbidity and would enable future research on this important question. My aim is therefore to study how DNAm age is expressed in schizophrenia and if DNAm aging, by measures of epigenetic age acceleration and age deceleration, is impacted in patients diagnosed with the illness. At the start of the project, multiple functional aging predictors had been studied, but no clear demonstration of differential aging in schizophrenia was reported (Nguyen, Eyler, and Jeste 2018). DNAm clocks offer a new perspective to study aging in schizophrenia as markers of biological age and predictors of mortality. I first discuss a methodological study I conducted on DNAm predictors' sensitivity for technical variation in general and the prospects for implementing these genomic tools in the clinic. I then discuss the outcomes of an in-depth analysis of the DNAm aging landscape in schizophrenia, and the implications of my findings for research and clinical utilization.

Take-home 5: DNAm-based predictors are promising new genomic tools but require careful implementation

Before discussing how DNAm age is expressed in schizophrenia, I will first discuss the reliability of DNAm-based predictors and their potential for clinical utilization. Both DNAm clocks and DNAm-based predictors hold great promise as genomic tools for health interventions and disease management in the clinic (McCartney et al. 2018; Shah et al. 2015). However, like other high-throughput molecular data, DNAm can be impacted by variations in sample handling, laboratory conditions, reagents, and/or equipment used (Leek et al. 2010). Technical variation is often widespread and tackling such effects is of critical importance to study biological variation in any -omic analysis, including DNAm. To investigate how reliable DNAm-based predictors are, I performed a systematic evaluation of the performance of 41 predictors, including multiple DNAm clocks, across more than 100 commonly used data processing and normalization strategies. These analytical strategies represent different ways to prepare DNAm data using published methodology, which I will refer to as "pipelines". In this study, I made use of a large sample of technical EPIC array DNAm replicates collected by the JacksonHeart Study. This allowed me to quantify the average absolute agreement between replicate pairs as a measure of test-retest reliability for each predictor. The test-retest reliability is an index of internal validity and indicates to what degree a method can produce outcomes that are reproducible (Koo and Li 2016). As high test-retest reliability is a necessary criterion for a method to qualify as a biomarker in clinical medicine, my work has important value for both research purposes and clinical utilization. This study was an enormous undertaking and represents the largest systematic and unbiased evaluation of the performance of DNAm-based algorithms.

I found that the performance of DNAm-based predictors is highly sensitive to technical variation, and that the choice of analytical pipeline to prepare DNAm data has a significant impact on their reliability and downstream phenotypic analyses. Having said that, all predictors produced at least good to excellent test-retest reliability if an appropriate analytical pipeline is used. The majority of predictors (32 out of 41) even achieved excellent reliability, indicating that estimates of these predictors are highly reproducible if unwanted technical variation is successfully removed. Out of the 41 best performing pipelines, 27 were unique to a predictor, demonstrating that significant heterogeneity exists in pipeline performance across predictors. There is thus no one-size fits all approach for analyzing these methods. Data processing and normalization strategies from the ENmix software package performed best most often. Data processing steps that performed well for multiple predictors were out-of-band (OOB) background estimation (Triche et al. 2013), REgression on Logarithm of Internal Control (RELIC) probes dye-bias correction (Xu et al. 2017), and the Regression on Correlated Probes (RCP) probe-type bias correction (Xu et al. 2017; Niu, Xu, and Taylor 2016). These methods leverage specific features of the EPIC DNAm array, such as internal control probes and the distribution of type I and II probes, to account for technical variation. While such analytical strategies performed well for most predictors, their performance still varied considerably. I therefore provided data processing and normalization best practices for each predictor for the research community to use. The programming code for all the analytical pipelines are available too, so they can be implemented by others easily. My aim is that this will help improve the performance of the algorithms and increase comparability of results across studies by standardization of data processing and normalization pipelines implemented by the research community.

Several predictors show a test-retest reliability close to 1, which means that repeated collections of DNAm data yielded almost identical predictor estimates between pairs of replicates. Among predictors with excellent reliability (intraclass correlation > 0.90) are predictors of ageing, mortality risk, smoking behavior, blood cell types, plasma protein levels, and cancer risk. Their high reproducibility makes them strong candidates for clinical biomarkers to aid in health management and disease prevention. GrimAge, a strong predictor of all-cause mortality, for example, showed the highest test-retest reliability. A one-year increase in predicted GrimAge compared to chronological age, gives an individual a 10% higher risk of dying (Lu et al. 2019). Implementation of GrimAge as a biomarker in the clinic could, therefore, have a significant impact on health and lifespan. While GrimAge had excellent reliability across all 101 pipelines (reliability range = 0.921-0.994), the performance of a pipeline still significantly impacted downstream phenotypic analyses. Pipelines that more effectively remove unwanted technical variation yielded a weaker correlation with chronological age, a lower mortality risk estimate, and greater statistical power in survival analyses than pipelines that were less successful in their performance. The choice of pipeline did not only impact downstream analysis of the GrimAge predictor. In fact, the distribution of output estimates of 80% of predictors showed changes in either the mean or standard deviation in relation to the performance of a pipeline, even when reliability of a predictor was high across most pipelines. This indicates that pipelines that produce improvements in DNAm-based predictor reliability, even if these are incremental improvements, can have significant impact on downstream phenotypic analyses. These findings therefore warrant the careful consideration of the choice of analytic pipeline when preparing DNAm data for implementation of these predictors.

My findings demonstrated good to excellent test-retest reliability for these 41 DNAm-based predictors based on technical replicate pairs that originated from the same biological sample. That is, a replicate pair represents a single DNA sample for which DNAm data was collected at two separate occasions. This indicates that DNAm-based predictors are promising candidates for biomarkers in the clinic based on their high reproducibility. It remains an open question if the measured reliability translates to repeated measures of DNA samples extracted from multiple blood draws at the same time point or across time points. This should be an important next step for investigation in future research. The analytical framework I applied can be easily extended to study designs of other types of (biological) replicates and establish method reliability in other contexts of technical and biological variation and across different studies. As research on DNAm-based predictors will continue to grow, my work produced new insights into algorithm performance and provides a comprehensive overview of best analytical practices when implementing these predictors for the research community to build on.

Take-home 6: DNA methylation age is affected in schizophrenia with age- and sex-specific effects

To investigate the expression of DNAm age in schizophrenia, I conducted a meta-analysis of three DNAm clocks (i.e., the Hannum, Horvath, and Levine clock) in a large multi-cohort case-control sample. The aim of this study was to investigate if DNA methylation age is different in individuals diagnosed with schizophrenia compared to non-psychiatric controls. In my analyses, I included three DNAm clocks that were developed in training datasets with specific biological characteristics

that allows them to capture different aspects of the ageing process. What I found was that DNAm age was significantly altered in schizophrenia with age- and sex-specific effects for multiple clocks. Furthermore, I observed an intriguing association between DNAm age acceleration and schizophrenia polygenic risk. This represents the largest study of DNAm age in schizophrenia and highlights DNAm clocks as novel genomic tools for quantifying changes in biological age in relation to the illness. The strengths of the study are its meta-analytical framework and the careful dissection of DNAm aging effects by age and sex. A systematic review of aging biomarkers found that less than a quarter of studies performed age-stratified analyses (Nguyen, Eyler, and Jeste 2018). My findings support their recommendation to specifically examine interaction effects with age and sex in aging studies, but also more general in epigenetic studies, for example like epigenome-wide association studies. Finally, my analyses are also the first to integrate DNAm age with genetic data and detailed phenotypes, such as age of onset and illness duration in a schizophrenia population. While it remains unclear whether the observed differences in ageing are intrinsic to schizophrenia or if they represent molecular consequences of the illness, my findings present an in-depth analysis of DNAm age in schizophrenia and unequivocally demonstrate that biological age is altered.

One of the main findings of the study is that people with a schizophrenia diagnosis present an increase of +1.53 years in Levine DNAm age compared to their chronological age compared to controls. The observed Levine DNAm age acceleration (i.e., Levine Δage) is more pronounced in older adult women diagnosed with schizophrenia and independently contributes to the variance in disease status in these women above and beyond smoking scores and blood cell type proportions. A recent DNAm aging study replicated my finding of Levine DNAm age acceleration in schizophrenia. In a smaller cohort, with partially overlapping samples, researchers reported a +1.40 to +1.90 increase in Levine Δage in cases compared to controls (Higgins-Chen et al. 2020), which is similar to the +1.53 years increase in Levine DNAm age observed in my analysis. However, they did not perform age- or sex-stratified analyses, nor did they integrate DNAm age with schizophrenia polygenic risk. I found that Levine Δage is positively associated with schizophrenia polygenic risk in women in later adulthood, indicating that women with higher burdens of genetic risk of the illness display even faster age acceleration. The high polygenic risk group displays accelerated aging of an average of +4.30 years compared to age-matched female controls. The Levine clock is constructed by predicting a surrogate measure of phenotypic age (also called PhenoAge), which is a weighted average of ten clinical markers known to be associated with mortality risk, including chronological age, albumin, creatinine, glucose and C-reactive protein levels, alkaline phosphatase, and various blood cell related measures (Levine et al. 2018). By design, the Levine estimator is a composite biomarker that strongly predicts mortality, in particular that of cardiovascular-related phenotypes. A one-year increase in Levine DNAm age is associated with a 9% increased risk of all-cause mortality and a 10% and 20% increase of cardiovascular disease and diabetes mortality risk respectively (Levine et al. 2018). My findings of multiple year increase in Levine DNAm age in schizophrenia could thus imply an increased mortality in these individuals that is linked to disease.

In contrast to Levine DNAm age acceleration, I observed age deceleration in Horvath DNAm age, which measures intrinsic cellular age. Horvath Δage showed strong age-specific effects but no clear sex-specific effects. During young adulthood, cases were -1.23 years younger in Horvath DNAm age compared to their chronological age than controls. Cases also presented Horvath DNAm age deceleration during their mid-forties. Horvath DNAm aging has been shown to be associated with molecular processes of development and cell differentiation, including human (neuro)developmental phenotype. For example, individuals diagnosed with Tatton-Brown Rahman syndrome and Sotos syndrome, two overgrowth conditions, present accelerated Horvath DNAm age, while people with Kabuki syndrome, clinically characterized by poor growth, display decelerated Horvath DNAm age (Jeffries et al. 2019). The observation of Horvath DNAm age deceleration in schizophrenia stands out because the majority of DNAm age studies of health-related outcomes report evidence of epigenetic age acceleration, but rarely of epigenetic age deceleration (Oblak et al. 2021). For schizophrenia, it may indicate that some patients show evidence of delayed or deficient development, and that this is detectable in blood through the multi-tissue Horvath clock. While speculative, this does present an intriguing hypothesis for future research. Horvath DNAm age deceleration in schizophrenia is mostly observed during developmentally sensitive windows, i.e., during late adolescence/young adulthood and the age-period of menopause. The etiology of schizophrenia has, furthermore, been shown to overlap with specific neurodevelopmental disorders (Owen et al. 2011). I did not observe age deceleration in postmortem brain samples of the human cortex, suggesting that the observed effect of epigenetic age deceleration is blood-specific. The overall cohort sample size of the postmortem brain analysis was, however, small and analyses in larger cohorts are therefore warranted. There was no heterogeneity in the effect size across the different cohorts used in the meta-analysis in blood tissue, making the finding of DNAm age deceleration in schizophrenia more robust and less likely to be a false positive. How blood-based Horvath DNAm age deceleration is associated with the trajectory of the illness and clinical outcomes and if it is at all related to a delayed or deficient development remains to be deciphered.

One fascinating outcome of my epigenetic investigations in schizophrenia is the opposite aging effect with two different DNAm clocks, i.e., DNAm age deceleration with the Horvath clock and age acceleration with the Levine clock. I did not observe differential DNAm aging for the Hannum clock. Unlike the multi-tissue Horvath clock, which was developed across different human tissues and cell types (Horvath 2013), the Hannum and Levine clocks were developed using DNAm samples derived from whole blood (Hannum et al. 2013). As the Horvath clock was trained using a wide variety of different cell types, it captures aging processes intrinsic to cells, while the Levine clock (and the Hannum clock) captures aging processes extrinsic to cells. Intrinsic cellular aging is less dependent on cell type composition and captures aging processes shared across cell types (Horvath and Raj 2018). My finding of opposite differential effects in intrinsic and extrinsic cellular aging processes suggests that two different aspects of biological age may be affected in schizophrenia. It has been hypothesized that schizophrenia is a syndrome of accelerated aging (Kirkpatrick et al. 2008; Nguyen, Eyler, and Jeste 2018), but my findings indicate that biological age is affected more broadly in schizophrenia, and that the precise mechanisms are more complex.

This is in line with evidence from studies of brain age in schizophrenia. A longitudinal brain age study using structural MRI scans for example reported age accelerated in the brain (Schnack et al., n.d.). They found that patterns of differential brain aging were not constant and suggested the involvement of two different aging processes; one that is homogeneous and reflective of brain age acceleration and another one that is reflective of individual variation and possibly medication use. In a study that I co-authored, we investigated if blood-based DNAm age and brain-based MRI age are associated. What we found was that the accelerated aging observed in the brain and blood reflects distinct biological processes (Teeuw et al. 2021). While this represents a first explorative study of brain and blood aging in a small cohort and requires replication in a larger sample, it implies that tissues are impacted differentially. In a large collaborative effort, the ENIGMA Epigenetics Working Group, which I was involved in as well, demonstrated that blood-based DNAm CpG sites are predictive of brain volume of the hippocampus in a large cohort (Jia et al. 2019). Future studies that aim to link blood and brain aging in larger samples could therefore be a worthwhile effort, as our current analyses may have been underpowered. Combining blood-based DNAm age with that of different bodily ages and from multiple cellular levels could better capture various aspects of biological aging and thereby advance our understanding of aging in schizophrenia (Cole et al. 2018; Jansen et al. 2021).

Future research on DNAm age in schizophrenia ideally would be conducted in prospective cohorts with detailed records on disease onset and progression of patients and medication use available. A limitation of my work is the cross-sectional design of the cohorts used and the lack of information on medication use, which probably affected the observed outcomes to some extent. While I did find an association between DNAm age acceleration and schizophrenia polygenic risk, dissecting cause-and-effect relationships was not possible. While one would, of course, ideally use a longitudinal study design in future research, cohorts that can enable such analyses are unfortunately rare. Developing new statistical methodology to extract causal relationships based on cross-sectional study design, by studying families or by extending mendelian randomization analyses to DNAm data, for example, should be explored as well. Taken together, my work demonstrates that DNAm age is affected in schizophrenia, a population vulnerable to age-related diseases and excess mortality, and strengthens the need for more research on the role of blood-based DNAm age in schizophrenia.

Take-home 7: Blood-based DNAm predictors hold promise as clinical tools for schizophrenia, but more research and funding are needed

Excess mortality in schizophrenia is a well-established epidemiological observation (Olfson et al. 2015; McGrath et al. 2008). My finding of significant Levine DNAm age acceleration suggests that the increased mortality risk could be quantified at the molecular level. Early detection of those with a predicted shorter lifespan has potential for high clinical impact, particularly for people diagnosed with schizophrenia who, on average, live 15 years shorter compared to the general population. The Levine clock not only predicts all-cause mortality but is also a strong predictor of cardiovascular-related mortality. People who suffer from schizophrenia have a two- to three-fold increase in cardiovascular-related mortality risk as well, women in particular die disproportionately of cardiovascular diseases (Olfson et al. 2015; Saha, Chant, and McGrath

2007). Adverse health outcomes associated with weight gain and/or metabolic syndrome as a consequence of antipsychotic medication are believed to be a main contributor to the observed excess mortality (Remington 2006). The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia study estimated that men are 138% more likely to have metabolic syndrome than randomly matched controls, and women 251% (McEvoy et al. 2005). Metabolic syndrome and its accompanying diseases are thus highly prevalent in patients and represent an enormous source of cardiovascular risk, especially for women. Minimizing metabolic risk is therefore of high priority for people diagnosed with schizophrenia, particularly those receiving long-term antipsychotic medication (Hert et al. 2009). As cardiovascular risk is modifiable and cardioprotective medication in people taking antipsychotic medication has been shown to reduce mortality risk (Kugathasan et al. 2018), the Levine DNAm clock could serve as a potential biomarker to identify at-risk individuals and, in this way, help with disease management and the improvement of life-expectancy in people with schizophrenia. The Levine DNAm clock reported a test-retest reliability of 0.97 in my analyses of algorithm performance and thus makes for an excellent candidate for further research. One of my main findings is that women in later adulthood with a diagnosis of schizophrenia showed the fastest Levine DNAm age acceleration. Women are overrepresented among those who develop the illness in later adulthood and are at higher risk for neglect of medical care (Dickerson 2007). My findings warrant a more focused and larger study of DNAm aging in women in later adulthood, preferably stratified by polygenic risk of the illness.

Reducing the burden of age-related disabilities and morbidities is an important goal in medicine. The Global Burden of Disease Study identified that age-related diseases account for more than half of the total burden of disease in 2017 (Chang et al. 2019). DNAm-based predictors could serve as quantitative biomarkers for health care prevention and disease management, particularly for psychiatric disorders. Identification of biomarkers for psychiatric care is one of the grand challenges in global mental health (Collins et al. 2011). In this dissertation, I have described why the study of ageing is important in schizophrenia and how biological age is affected in detail. While my findings are promising, more research is needed to replicate results and investigate the degree of clinical actionability that can be achieved when using DNAm clocks or other DNAm-based predictors. In a first attempt to assess the clinical value of DNAm clocks, I co-authored a study on the association between DNAm age acceleration and all-cause mortality in a Swedish cohort of 190 schizophrenia cases and 190 controls, with a 2:1 oversampling of individuals who died (Kowalec et al. 2019). We investigated the Hannum, Horvath, and Levine clock, but did not find a significant association between DNAm age acceleration and mortality in schizophrenia for any of the clocks. Despite the small cohort, we had >80% power to detect a hazard ratio of 1.17 given the study design and sample size. This warrants caution and emphasizes the need for more research on the role of DNAm age in schizophrenia and its predictive capacity for mortality risk. The sample that was used in the Swedish study did include more men than women. In fact, 76% of the cases and 71% of the controls who died were male, which likely decreased statistical power in our analyses. Given my finding of stronger Levine DNAm age acceleration in women diagnosed with schizophrenia, future studies on the association between DNAm age acceleration and mortality in schizophrenia should prioritize the inclusion of women.

Future research should also consider the development of schizophrenia specific ageing

biomarkers, for example, by developing DNAm clocks with sensitivity to medication use or for other phenotypic characteristics related to the trajectory of the illness, like substance use or the degree of chronicity. Combining DNA methylation clocks with other types of biological data is an important avenue to explore as well. A recent study that combined five - omic based biological clocks (telomere length, epigenetic, transcriptomic, proteomic, and metabolomic clocks) showed a low correlation among predictors and demonstrated that a composite index that combined all five blood-based clocks yielded the strongest association with various health outcomes (Jansen et al. 2021). This suggests that one's biological age is best reflected by combining aging measures from multiple data layers. As awareness of the value of peripheral biomarkers of schizophrenia is increasing (Lai et al. 2016), we need more research on and allocation of funding towards blood-based -omic analyses in schizophrenia. The figure below visualizes the number of publications on schizophrenia genetic and epigenetic research over time (panel A) and highlights how blood-based epigenetic studies are still lagging behind that of genetic studies.

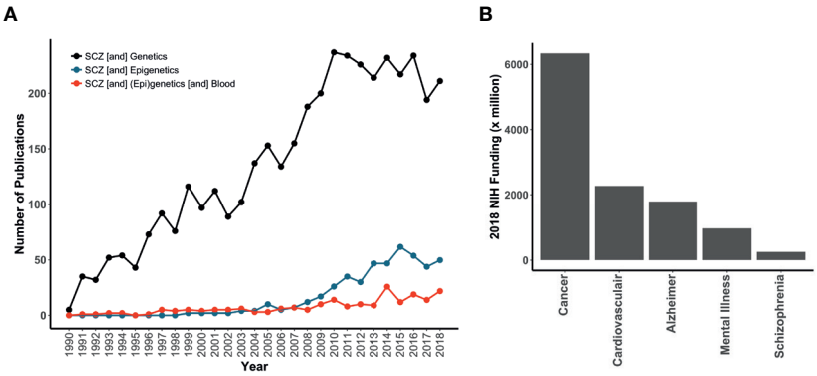


Figure 2. Blood-based epigenetic research is lagging behind in schizophrenia, as is research funding. *(A) Shown are the number of publications per year between 1990 and 2018 for schizophrenia (epi)genetic research. Web of Science was used with the following queries; (1 - black) TS=("schizophrenia" AND "genetics"), (2 - blue) TS=("schizophrenia" AND "epigenetics"), (3 - red) TS=("schizophrenia" AND "blood" AND ("genetics" OR "epigenetics"). Terms were searched for in abstract, title, and keyword fields of a publication. (B) Overall funding of the National Institute of Health (NIH) in 2018 shown for specific disease groups.*

Collecting genome-wide -omic data, including DNAm data, is expensive and research on schizophrenia is underfunded compared to other illnesses (panel B), despite its high disease burden. Therefore, I believe that expanding resources to perform blood-based epigenetic research and other blood-based -omic research will be a critical component of translating the value of ageing and health predictors towards clinical actionability for schizophrenia. For now, it remains an open question whether DNA methylation clocks are the much-needed biomarker to aid in the care of people diagnosed with schizophrenia.

Take-home 8: Open and collaborative science is the path forward

I was fortunate to collaborate with many national and international colleagues in my research efforts. I have, for example, been involved in working groups of the Psychiatric Genomics Consortium (PGC) and contributed as a data analyst to research projects that involved hundreds of colleagues across the world (Stahl et al. 2019; Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2018). This has shaped my vision of science and what can be achieved when we operate as a collective. In addition, many key data analyses in my research made use of publicly available data and software generated and developed by others. These were essential components that helped me conduct my research and shaped the outcome of this dissertation. In the functional investigations of schizophrenia polygenic risk, for example, I used the GWAS summary statistics of genetic analyses of hundreds of thousand individuals that were made available through the PGC (Sullivan et al. 2017) to map heritability to *in vitro* transcriptomic profiles. Furthermore, I replicated the enrichment of schizophrenia heritability I observed in differentially expressed genes during neuronal differentiation in the CORTECON RNA sequencing resource of human cortical development (van de Leemput et al. 2014), which I freely downloaded from the GEO database to use as an independent replication dataset. In the same study, I also mapped *in vitro* transcriptomic profiles associated with neuronal differentiation to spatiotemporal dimension of human brain development using gene expression of the Allen Brain Atlas resource. The Allen Human Brain Atlas is a freely available multimodal atlas of gene expression and anatomy consisting of a detailed overview of molecular and cellular data of the human brain (Shen, Overly, and Jones 2012). In my study of clozapine response in lymphoblast cell lines, I investigate how the identified clozapine-associated genes were expressed in the GTEx gene expression dataset of 22 human tissues. The GTEx data, freely available for download from the GTEx web portal and part of the GTEx biobank, represents the most comprehensive atlas of human gene expression across tissues (Lonsdale et al. 2013). By using data from the Allen Human Brain Atlas and the GTEx project, I was able to investigate how the *in vitro* gene expression profiles observed in my experiments translated to human biology across tissues and developmental periods. The wealth of publicly available human genomic data represents a treasure trove for new research and *in vitro* experimental studies need to leverage these datasets more. I hope my research will contribute to this by leading by example.

Similar to my *in vitro* functional investigations, open access genomic data was important in my analyses of DNAm age as well. That is, I included DNAm data of a British and a Scottish cohort in the meta-analysis of DNAm aging in schizophrenia alongside DNAm data from several Dutch cohorts. The former two datasets are available through the Gene Expression Omnibus (GEO) database and have been used in large schizophrenia DNAm studies in the past (Hannon et al. 2016). In the same meta-analysis, a Swedish case-control cohort was also included via a collaboration with colleagues at the University of North Carolina at Chapel Hill in the USA and the Karolinska Institutet in Stockholm, Sweden. From the British and Scottish cohorts, phenotype data on age of onset and illness duration of cases, alongside schizophrenia polygenic risk scores, were made available by principal investigators of the cohorts and could therefore be analyzed jointly with similar data of the Dutch cohort. The sharing of both genetic and phenotype data across studies is what made the in-depth analyses of DNAm aging landscape in schizophrenia

possible. Finally, in my investigating of the test-retest reliability of DNAm-based predictors, I used DNAm data from the JacksonHeart Study, a large community-based cohort study evaluating the etiology of cardiovascular, renal, and respiratory diseases among African Americans in the Jackson, Mississippi metropolitan area in the USA (Taylor et al. 2005). They offer access to data collected by the cohort. Their large number of EPIC DNAm array technical replicate pairs were essential to evaluating the impact of technical variation on the performance of DNAm-based predictors.

In the output of my research, I have tried to pay the opportunities that open and collaborative science provided me forward. De-identified genomic data that was collected as part of the studies in this dissertation were uploaded to the GEO database and all manuscripts posted on bioRxiv, an open access preprint repository for biological sciences. The programming code to run analyses of the 101 data processing and normalization pipelines of DNAm data was uploaded to my github repository. Through these open science practices, I aim to make the dissemination of my research output accessible to all levels and stakeholders in society. Perhaps the most valuable lessons I have learned during my academic journey so far are the importance of a strong commitment to making information and knowledge transparent and accessible, and what can be achieved through collaborative science.

Conclusion

This dissertation had the ambitious aim to disentangle and, if possible, to clarify some of the biological complexity that underlies schizophrenia, a debilitating illness that affects millions of people worldwide. While little is known about the precise molecular mechanisms that cause a person to fall ill or that relate to consequences of the illness, genomic research is accelerating our understanding of the biology of schizophrenia. This dissertation embodies the genomic opportunities and diverse research strategies that lay at our disposal to clarify the biomedical complexity of schizophrenia. I have summarized the output of my work in eight important take-home messages.

I demonstrated the usefulness of laboratory experimental systems to study schizophrenia biology and the importance of embracing polygenicity in functional investigation of its genetic risk. I highlighted that schizophrenia polygenic risk maps to synaptic biology and that *in vitro* human neuronal differentiation can serve as a valuable model system to further map schizophrenia heritability in the post-GWAS era. Similarly, I described lymphoblast cell lines as a useful experimental tool for studying the biology of clozapine response and that more research is needed to understand the therapeutic mechanisms and molecular adverse effects of this effective antipsychotic drug. I demonstrated the importance of accounting for technical variation when analyzing DNAm-based predictors and how these algorithms have good to excellent test-rest reliability making them suitable biomarker candidates. This could have implications for schizophrenia, as I showed that Levine DNAm age is affected in the illness, and that this effect is particularly strong in women in later adulthood. I plead for more research and funding allocation towards blood-based DNAm investigations as the study of DNAm clocks can have

high downstream clinical impact for schizophrenia. Finally, I emphasized the importance of open and collaborative science, without which the research in this dissertation would not have been possible.

Taken together, these functional investigations represent in-depth data-driven efforts to gain new knowledge of schizophrenia biology through integrative genomic analyses. My findings provide important building blocks for future genomic research to understand the molecular causes and consequences of the illness. While there are still many challenges to overcome, my research highlights that the avenues to do so have never been more open.

References

Allebeck, P. 1989. "Schizophrenia: A Life-Shortening Disease." Schizophrenia Bulletin 15 (1): 81–89.

Andersohn, Frank, Christine Konzen, and Edeltraut Garbe. 2007. "Systematic Review: Agranulocytosis Induced by Nonchemotherapy Drugs." Annals of Internal Medicine 146 (9): 657–65.

Bablenis, E., S. S. Weber, and R. L. Wagner. 1989. "Clozapine: A Novel Antipsychotic Agent." DICP: The Annals of Pharmacotherapy 23 (2): 109–15.

Bai, C., R. Richman, and S. J. Elledge. 1994. "Human Cyclin F." The EMBO Journal 13 (24): 6087–98.

Banker, G. A. 1980. "Trophic Interactions between Astroglial Cells and Hippocampal Neurons in Culture." Science 209 (4458): 809–10.

Baptista, T. 1999. "Body Weight Gain Induced by Antipsychotic Drugs: Mechanisms and Management." Acta Psychiatrica Scandinavica 100 (1): 3–16.

Bardy, Cedric, Mark van den Hurk, Tameji Eames, Cynthia Marchand, Ruben V. Hernandez, Mariko Kellogg, Mark Gorris, et al. 2015. "Neuronal Medium That Supports Basic Synaptic Functions and Activity of Human Neurons *in Vitro*." Proceedings of the National Academy of Sciences of the United States of America 112 (20): E2725–34.

Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address: douglas.ruderfer@vanderbilt.edu, and Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2018. "Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes." Cell 173 (7): 1705–15.e16.

Buxbaum, Joseph, Kristen Brennand, Andrew Browne, Elodie Drapeau, Katie Brenner, Scott Noggle, and Rachel Yehuda. 2019. "Large-Scale Reprogramming and Neuronal Differentiation In Complex Psychiatric Disorders." European Neuropsychopharmacology.

Chang, Angela Y., Vegard F. Skirbekk, Stefanos Tyrovolas, Nicholas J. Kassebaum, and Joseph L. Dieleman. 2019. "Measuring Population Ageing: An Analysis of the Global Burden of Disease Study 2017." The Lancet. Public Health 4 (3): e159–67.

Chaudhury, Sraboni, Vikram Sharma, Vivek Kumar, Tapas C. Nag, and Shashi Wadhwa. 2016. "Activity-Dependent Synaptic Plasticity Modulates the Critical Phase of Brain Development." Brain & Development 38 (4): 355–63.

Cole, James H., Riccardo E. Marioni, Sarah E. Harris, and Ian J. Deary. 2018. "Brain Age and Other Bodily 'Ages': Implications for Neuropsychiatry." Molecular Psychiatry, June.

Collins, Pamela Y., Vikram Patel, Sarah S. Joestl, Dana March, Thomas R. Insel, Abdallah S. Daar, Scientific Advisory Board and the Executive Committee of the Grand Challenges on Global Mental Health, et al. 2011. "Grand Challenges in Global Mental Health." Nature 475 (7354): 27–30.

Consortium, Schizophrenia Working Group of The Psychiatric Genomics, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Stephan Ripke, James T. R. Walters, and Michael C. O'Donovan. n.d. "Mapping Genomic Loci Prioritises Genes and Implicates Synaptic Biology in Schizophrenia."

Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu, and Cross-Disorder Group of the Psychiatric Genomics Consortium. 2019. "Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders." Cell 179 (7): 1469–82.e11.

Das, Debamitra, Xi Peng, Anh-Thu N. Lam, Joel S. Bader, and Dimitrios Avramopoulos. 2021. "Transcriptome Analysis of Human Induced Excitatory Neurons Supports a Strong Effect of Clozapine on Cholesterol Biosynthesis." Schizophrenia Research 228 (February): 324–26.

Dibonaventura, Marco, Susan Gabriel, Leon Dupclay, Shaloo Gupta, and Edward Kim. 2012. "A Patient Perspective of the Impact of Medication Side Effects on Adherence: Results of a Cross-Sectional Nationwide Survey of Patients with Schizophrenia." BMC Psychiatry 12 (March): 20.

Dickerson, Faith B. 2007. "Women, Aging, and Schizophrenia." Journal of Women & Aging. Fass, D., S. Blacklow, P. S. Kim, and J. M. Berger. 1997. "Molecular Basis of Familial Hypercholesterolaemia from Structure of LDL Receptor Module." Nature 388 (6643): 691–93.

Fernández, Esperanza, Nicholas Rajan, and Claudia Bagni. 2013. "The FMRP Regulon: From Targets to Disease Convergence." Frontiers in Neuroscience 7 (October): 191.

Fernø, J., M. B. Raeder, A. O. Vik-Mo, S. Skrede, M. Glambek, K-J Tronstad, H. Breilid, et al. 2005. "Antipsychotic Drugs Activate SREBP-Regulated Expression of Lipid Biosynthetic Genes in Cultured Human Glioma Cells: A Novel Mechanism of Action?" The Pharmacogenomics Journal 5 (5): 298–304.

Finucane, Hilary K., Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. "Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types." Nature Genetics 50 (4): 621–29.

Foley, D. L., and A. Mackinnon. 2014. "A Systematic Review of Antipsychotic Drug Effects on Human Gene Expression Related to Risk Factors for Cardiovascular Disease." The Pharmacogenomics Journal 14 (5): 446–51.

Fromer, Menachem, Andrew J. Pocklington, David H. Kavanagh, Hywel J. Williams, Sarah Dwyer, Padhraig Gormley, Lyudmila Georgieva, et al. 2014. "De Novo Mutations in Schizophrenia Implicate Synaptic Networks." Nature 506 (7487): 179–84.

Gallagher, Michael D., and Alice S. Chen-Plotkin. 2018. "The Post-GWAS Era: From Association to Function." American Journal of Human Genetics 102 (5): 717–30.

Galletly, Cherrie A., Debra L. Foley, Anna Waterreus, Gerald F. Watts, David J. Castle, John J. McGrath, Andrew Mackinnon, and Vera A. Morgan. 2012. "Cardiometabolic Risk Factors in People with Psychotic Disorders: The Second Australian National Survey of Psychosis." The Australian and New Zealand Journal of Psychiatry 46 (8): 753–61.

Gaspar, H. A., and G. Breen. 2017. "Drug Enrichment and Discovery from Schizophrenia Genome-Wide Association Results: An Analysis and Visualisation Approach." Scientific Reports 7 (1): 12460.

Genovese, Giulio, Menachem Fromer, Eli A. Stahl, Douglas M. Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L. Moran, et al. 2016. "Increased Burden of Ultra-Rare Protein-Altering Variants among 4,877 Individuals with Schizophrenia." Nature Neuroscience, no. October

Glessner, Joseph T., Muredach P. Reilly, Cecilia E. Kim, Nagahide Takahashi, Anthony Albano, Cuiping Hou, Jonathan P. Bradfield, et al. 2010. "Strong Synaptic Transmission Impact by Copy Number Variations in Schizophrenia." Proceedings of the National Academy of Sciences of the United States of America 107 (23): 10584–89.

Goff, Donald C., Peter Falkai, W. Wolfgang Fleischhacker, Ragy R. Girgis, Rene M. Kahn, Hiroyuki Uchida, Jingping Zhao, and Jeffrey A. Lieberman. 2017. "The Long-Term Effects of Antipsychotic Medication on Clinical Course in Schizophrenia." The American Journal of Psychiatry 174 (9): 840–49.

Gonçalves, Pedro, João Ricardo Araújo, and Fátima Martel. 2015. "Antipsychotics-Induced Metabolic Alterations: Focus on Adipose Tissue and Molecular Mechanisms." European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology 25 (1): 1–16.

Grotzinger, Andrew D., Mijke Rhemtulla, Ronald de Vlaming, Stuart J. Ritchie, Travis T. Mallard, W. David Hill, Hill F. Ip, et al. 2019. "Genomic Structural Equation Modelling Provides Insights into the Multivariate Genetic Architecture of Complex Traits." Nature Human Behaviour.

Hall, Jeremy, Simon Trent, Kerrie L. Thomas, Michael C. O'Donovan, and Michael J. Owen. 2015. "Genetic Risk for Schizophrenia: Convergence on Synaptic Pathways Involved in Plasticity." Biological Psychiatry.

Hannon, Eilis, Emma Dempster, Joana Viana, Joe Burrage, Adam R. Smith, Ruby Macdonald, David St Clair, et al. 2016. "An Integrated Genetic-Epigenetic Analysis of Schizophrenia: Evidence for Co-Localization of Genetic Associations and Differential DNA Methylation." Genome Biology 17 (1): 176.

Hannon, Eilis, Olivia Knox, Karen Sugden, Joe Burrage, Chloe C. Y. Wong, Daniel W. Belsky, David L. Corcoran, et al. 2018. "Characterizing Genetic and Environmental Influences on Variable DNA Methylation Using Monozygotic and Dizygotic Twins." PLoS Genetics 14 (8): e1007544.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srinivas Sadda, Brandy Klotzle, et al. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." Molecular Cell 49 (2): 359–67.

Harrison, Rebecca N. S., Robin M. Murray, Sang Hyuck Lee, Jose Paya Cano, David Dempster, Charles J. Curtis, Danai Dima, Fiona Gaughran, Gerome Breen, and Simone de Jong. 2016. "Gene-Expression Analysis of Clozapine Treatment in Whole Blood of Patients with Psychosis." Psychiatric Genetics 26 (5): 211–17.

Hedegaard, Anne, Jimena Monzón-Sandoval, Sarah E. Newey, Emma S. Whiteley, Caleb Webber, and Colin J. Akerman. 2020. "Pro-Maturational Effects of Human iPSC-Derived Cortical Astrocytes upon iPSC-Derived Cortical Neurons." Stem Cell Reports 15 (1): 38–51. Hert, Marc D. E., Marc de Hert, Vincent Schreurs, Davy Vancampfort, and Ruud Van Winkel. 2009. "Metabolic Syndrome in People with Schizophrenia: A Review." World Psychiatry.

Higgins-Chen, Albert T., Marco P. Boks, Christiaan H. Vinkers, René S. Kahn, and Morgan E. Levine. 2020. "Schizophrenia and Epigenetic Aging Biomarkers: Increased Mortality, Reduced Cancer Risk, and Unique Clozapine Effects." Biological Psychiatry 88 (3): 224–35.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." Genome Biology 14 (10): R115.

Horvath, Steve, and Kenneth Raj. 2018. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing." Nature Reviews. Genetics 19 (6): 371–84.

Jaffe, Andrew E., Yuan Gao, Amy Deep-Soboslay, Ran Tao, Thomas M. Hyde, Daniel R. Weinberger, and Joel E. Kleinman. 2015. "Mapping DNA Methylation across Development, Genotype and Schizophrenia in the Human Frontal Cortex." Nature Neuroscience 19 (1): 40– 47.

Jansen, Rick, Laura Km Han, Josine E. Verhoeven, Karolina A. Aberg, Edwin Cgj van den Oord, Yuri Milaneschi, and Brenda Wjh Penninx. 2021. "An Integrative Study of Five Biological Clocks in Somatic and Mental Health." eLife 10 (February).

Jeffries, Aaron R., Reza Maroofian, Claire G. Salter, Barry A. Chioza, Harold E. Cross, Michael A. Patton, Emma Dempster, et al. 2019. "Growth Disrupting Mutations in Epigenetic Regulatory Molecules Are Associated with Abnormalities of Epigenetic Aging." Genome Research 29 (7): 1057–66.

Jia, Tianye, Congying Chu, Yun Liu, Jenny van Dongen, Evangelos Papastergios, Nicola J. Armstrong, Mark E. Bastin, et al. 2019. "Epigenome-Wide Meta-Analysis of Blood DNA Methylation and Its Association with Subcortical Volumes: Findings from the ENIGMA Epigenetics Working Group." Molecular Psychiatry, December.

Kirkpatrick, Brian, Erick Messias, Philip D. Harvey, Emilio Fernandez-Egea, and Christopher R. Bowie. 2008. "Is Schizophrenia a Syndrome of Accelerated Aging?" Schizophrenia Bulletin 34 (6): 1024–32.

Kirov, G., A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, et al. 2012. "De Novo CNV Analysis Implicates Specific Abnormalities of Postsynaptic Signalling Complexes in the Pathogenesis of Schizophrenia." Molecular Psychiatry 17 (2): 142–53.

Klein, Jason C., Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. 2020. "A Systematic Evaluation of the Design and Context Dependencies of Massively Parallel Reporter Assays." Nature Methods 17 (11): 1083–91.

Koopmans, Frank, Pim van Nierop, Maria Andres-Alonso, Andrea Byrnes, Tony Cijsouw, Marcelo P. Coba, L. Niels Cornelisse, et al. 2019. "SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse." Neuron 103 (2): 217–34.e4.

Koo, Terry K., and Mae Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." Journal of Chiropractic Medicine 15 (2): 155–63.

Kowalec, Kaarina, Eilis Hannon, Georgina Mansell, Joe Burrage, Anil P. S. Ori, Roel A. Ophoff, Jonathan Mill, and Patrick F. Sullivan. 2019. "Methylation Age Acceleration Does Not Predict Mortality in Schizophrenia." Translational Psychiatry.

Krebs, Catharine E., Anil P. S. Ori, Annabel Vreeker, Timothy Wu, Rita M. Cantor, Marco P. M. Boks, Rene S. Kahn, Loes M. Olde Loohuis, and Roel A. Ophoff. 2020. "Whole Blood Transcriptome Analysis in Bipolar Disorder Reveals Strong Lithium Effect." Psychological Medicine 50 (15): 2575–86.

Kugathasan, Pirathiv, Henriette Thisted Horsdal, Jørgen Aagaard, Svend Eggert Jensen, Thomas Munk Laursen, and René Ernst Nielsen. 2018. "Association of Secondary Preventive Cardiovascular Treatment After Myocardial Infarction With Mortality Among Patients With Schizophrenia." JAMA Psychiatry 75 (12): 1234–40.

Lai, Chi-Yu, Elizabeth Scarr, Madhara Udawela, Ian Everall, Wei J. Chen, and Brian Dean. 2016. "Biomarkers in Schizophrenia: A Focus on Blood Based Diagnostics and Theranostics." World Journal of Psychiatry 6 (1): 102–17.

Lee, Ji-Ann, Andrey Damianov, Chia-Ho Lin, Mariana Fontes, Neelroop N. Parikshak, Erik S. Anderson, Daniel H. Geschwind, Douglas L. Black, and Kelsey C. Martin. 2016. "Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes." Neuron 89 (1): 113–28.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael a. Irizarry. 2010. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." Nature Reviews. Genetics 11 (10): 733–39.

Leemput, Joyce van de, Nathan C. Boles, Thomas R. Kiehl, Barbara Corneo, Patty Lederman, Vilas Menon, Changkyu Lee, et al. 2014. "CORTECON: A Temporal Transcriptome Analysis of in Vitro Human Cerebral Cortex Development from Human Embryonic Stem Cells." Neuron 83 (1): 51–68.

Leucht, Stefan, Andrea Cipriani, Loukia Spineli, Dimitris Mavridis, Deniz Orey, Franziska Richter, Myrto Samara, et al. 2013. "Comparative Efficacy and Tolerability of 15 Antipsychotic Drugs in Schizophrenia: A Multiple-Treatments Meta-Analysis." The Lancet 382 (9896): 951– 62.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." Aging 10 (4): 573–91.

Lewis, Steven M., Adam Williams, and Stephanie C. Eisenbarth. 2019. "Structure and Function of the Immune System in the Spleen." Science Immunology 4 (33).

Li, Xiao-Hong, Xiao-Mei Zhong, Li Lu, Wei Zheng, Shi-Bin Wang, Wen-Wang Rao, Shuai Wang, et al. 2020. "The Prevalence of Agranulocytosis and Related Death in Clozapine-Treated Patients: A Comprehensive Meta-Analysis of Observational Studies." Psychological Medicine 50 (4): 583–94.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." Nature Genetics 45 (6): 580–85.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, et al. 2019. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan." Aging 11 (2): 303–27.

Malhotra, Dheeraj, Shane McCarthy, Jacob J. Michaelson, Vladimir Vacic, Katherine E. Burdick, Seungtai Yoon, Sven Cichon, et al. 2011. "High Frequencies of de Novo CNVs in Bipolar Disorder and Schizophrenia." Neuron 72 (6): 951–63.

McCartney, Daniel L., Robert F. Hillary, Anna J. Stevenson, Stuart J. Ritchie, Rosie M. Walker, Qian Zhang, Stewart W. Morris, et al. 2018. "Epigenetic Prediction of Complex Traits and Death." Genome Biology 19 (1): 136.

McEvoy, Joseph P., Jonathan M. Meyer, Donald C. Goff, Henry A. Nasrallah, Sonia M. Davis, Lisa Sullivan, Herbert Y. Meltzer, John Hsiao, T. Scott Stroup, and Jeffrey A. Lieberman. 2005. "Prevalence of the Metabolic Syndrome in Patients with Schizophrenia: Baseline Results from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Schizophrenia Trial and Comparison with National Estimates from NHANES III." Schizophrenia Research 80 (1): 19– 32.

McGrath, John, Sukanta Saha, David Chant, and Joy Welham. 2008. "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality." Epidemiologic Reviews 30 (May): 67–76.

Mijovic, Aleksandar, and James H. MacCabe. 2020. "Clozapine-Induced Agranulocytosis." Annals of Hematology 99 (11): 2477–82.

Mill, Jonathan, Thomas Tang, Zachary Kaminsky, Tarang Khare, Simin Yazdanpanah, Luigi Bouchard, Peixin Jia, et al. 2008. "Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis." American Journal of Human Genetics 82 (3): 696–711.

Nguyen, Tanya T., Lisa T. Eyler, and Dilip V. Jeste. 2018. "Systemic Biomarkers of Accelerated Aging in Schizophrenia: A Critical Review and Future Directions." Schizophrenia Bulletin 44 (2): 398–408.

Niu, Liang, Zongli Xu, and Jack A. Taylor. 2016. "RCP: A Novel Probe Design Bias Correction Method for Illumina Methylation BeadChip." Bioinformatics 32 (17): 2659–63.

Oblak, Lara, Jeroen van der Zaag, Albert T. Higgins-Chen, Morgan E. Levine, and Marco P. Boks. 2021. "A Systematic Review of Biological, Social and Environmental Factors Associated with Epigenetic Clock Acceleration." Ageing Research Reviews 69 (April): 101348.

Olde Loohuis, Loes M. Olde, Jacob A. S. Vorstman, Anil P. Ori, Kim A. Staats, Tina Wang, Alexander L. Richards, Ganna Leonenko, et al. 2015. "Genome-Wide Burden of Deleterious Coding Variants Increased in Schizophrenia." Nature Communications 6: 7501.

Olfson, Mark, Tobias Gerhard, Cecilia Huang, Stephen Crystal, and T. Scott Stroup. 2015. "Premature Mortality Among Adults With Schizophrenia in the United States." JAMA Psychiatry 72 (12): 1172–81.

Owen, Michael J., Michael C. O'Donovan, Anita Thapar, and Nicholas Craddock. 2011. "Neurodevelopmental Hypothesis of Schizophrenia." The British Journal of Psychiatry: The Journal of Mental Science 198 (3): 173.

Patwardhan, Rupali P., Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, et al. 2012. "Massively Parallel Functional Dissection of Mammalian Enhancers *in Vivo*." Nature Biotechnology 30 (3): 265–70.

Rebola, Nelson, Bettadapura N. Srikumar, and Christophe Mulle. 2010. "Activity-Dependent Synaptic Plasticity of NMDA Receptors." The Journal of Physiology.

Remington, Gary. 2006. "Schizophrenia, Antipsychotics, and the Metabolic Syndrome: Is There a Silver Lining?" American Journal of Psychiatry.

Saha, Sukanta, David Chant, and John McGrath. 2007. "A Systematic Review of Mortality in Schizophrenia: Is the Differential Mortality Gap Worsening over Time?" Archives of General Psychiatry 64 (10): 1123–31.
Schnack, Hugo G., Neeltje E. M. Van Haren, Mireille Nieuwenhuis, Hilleke E. Hulshoff Pol, Wiepke Cahn, Rudolf Magnus, and Centre Utrecht. n.d. "Accelerated Brain-Aging in Schizophrenia : A Longitudinal Pattern Recognition Study" 3420 (limit 3000).

Sekar, Aswin, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, et al. 2016. "Schizophrenia Risk from Complex Variation of Complement Component 4." Nature 530 (7589): 177–83.

Shah, Sonia, Marc J. Bonder, Riccardo E. Marioni, Zhihong Zhu, Allan F. McRae, Alexandra Zhernakova, Sarah E. Harris, et al. 2015. "Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations." American Journal of Human Genetics 97 (1): 75–85.

Shen, Elaine H., Caroline C. Overly, and Allan R. Jones. 2012. "The Allen Human Brain Atlas: Comprehensive Gene Expression Mapping of the Human Brain." Trends in Neurosciences 35 (12): 711–14.

Skene, Nathan G., Julien Bryois, Trygve E. Bakken, Gerome Breen, James J. Crowley, Héléna A. Gaspar, Paola Giusti-Rodriguez, et al. 2018. "Genetic Identification of Brain Cell Types Underlying Schizophrenia." Nature Genetics 50 (6): 825–33.

Stahl, Eli A., Gerome Breen, Andreas J. Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, et al. 2019. "Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder." Nature Genetics 51 (5): 793–803.

Südhof, Thomas C. 2018. "Towards an Understanding of Synapse Formation." Neuron.

Sullivan, Patrick F., Arpana Agrawal, Cynthia M. Bulik, Ole A. Andreassen, Anders D. Børglum, Gerome Breen, Sven Cichon, et al. 2017. "Psychiatric Genomics: An Update and an Agenda." The American Journal of Psychiatry, October, appiajp201717030283.

Taylor, Herman A., Jr, James G. Wilson, Daniel W. Jones, Daniel F. Sarpong, Asoka Srinivasan, Robert J. Garrison, Cheryl Nelson, and Sharon B. Wyatt. 2005. "Toward Resolution of Cardiovascular Health Disparities in African Americans: Design and Methods of the Jackson Heart Study." Ethnicity & Disease 15 (4 Suppl 6): S6–4 – 17.

Teeuw, Jalmar, Anil P. S. Ori, Rachel M. Brouwer, Sonja M. C. de Zwarte, Hugo G. Schnack, Hilleke E. Hulshoff Pol, and Roel A. Ophoff. 2021. "Accelerated Aging in the Brain, Epigenetic Aging in Blood, and Polygenic Risk for Schizophrenia." Schizophrenia Research 231 (April): 189–97.

Tragante, Vinicius, Daiane Hemerich, Mohammad Alshabeeb, Ingrid Brænne, Harri Lempiäinen, Riyaz S. Patel, Hester M. den Ruijter, et al. 2018. "Druggability of Coronary Artery Disease Risk Loci." Circulation. Genomic and Precision Medicine 11 (8): e001977.

Triche, Timothy J., Jr, Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. 2013. "Low-Level Processing of Illumina Infinium DNA Methylation BeadArrays." Nucleic Acids Research 41 (7): e90.

Vik-Mo, Audun O., Johan Fernø, Silje Skrede, and Vidar M. Steen. 2009. "Psychotropic Drugs up-Regulate the Expression of Cholesterol Transport Proteins Including ApoE in Cultured Human CNS- and Liver Cells." BMC Pharmacology 9 (August): 10.

With, S. A. J. de, S. L. Pulit, T. Wang, W. G. Staal, W. W. van Solinge, P. I. W. de Bakker, and R. A. Ophoff. 2015. "Genome-Wide Association Study of Lymphoblast Cell Viability after Clozapine Exposure." American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 168B (2): 116–22.

Xu, Zongli, Sabine A. S. Langie, Patrick De Boever, Jack A. Taylor, and Liang Niu. 2017. "RELIC: A Novel Dye-Bias Correction Method for Illumina Methylation BeadChip." BMC Genomics 18 (1): 4.

Appendices

Summary

One in every 200 people will be diagnosed with schizophrenia during their lifetime and about 1.5 million people received a schizophrenia diagnosis this year, worldwide. Despite the fact that so many people suffer from the illness and its high burden on families and societies, we know so little about the biology of schizophrenia with no cures yet available. Genomic technologies and new developments in quantitative genetic methodology have proven to be instrumental in advancing our knowledge on the biology of complex human diseases, including schizophrenia. Large-scale genetic studies have for example been successful at identifying genetic variation associated with schizophrenia thereby propelling new insights into its disease mechanisms. The next step forward is to conduct further functional investigations that build upon these genetic associations to clarify the molecular and cellular processes that are disrupted. At the same time, gaining understanding of the molecular consequences that impact the bodies and lives of those who suffer from the illness is equally important as well.

This thesis is a collection of my research on the biology of schizophrenia and embodies the genomic opportunities and diverse research strategies that lay at our disposal to clarify and improve our understanding of the illness. With the use of laboratory model systems of brain and blood cells, large case-control cohorts, newly developed DNA methylation algorithms, and by re-using and leveraging external genomic datasets, I conducted an in-depth investigation into the biology of schizophrenia. My aim was to go beyond the findings of large-scale genetic studies and conduct research that uses state-of-the-art methodology and integrative genomic analyses to identify new pieces to the puzzle of schizophrenia biology in the post-GWAS era. As an outcome of my research, I provide eight take-home messages that I believe are important conclusions for our understanding of the biology of schizophrenia or that are important insights on conducting psychiatric genomic research moving forward.

Clarifying the biological mechanisms associated with schizophrenia is a necessary step towards the development of new treatments and further humanization of this severe and stigmatized illness. My findings provide important building blocks for future genomic research to understand the molecular causes and consequences of the illness. While there are still many challenges to overcome, my research highlights that the avenues to do so have never been more open.

My eight important take-home messages:

1. We need to embrace polygenicity in functional investigations of schizophrenia.
2. Schizophrenia genetic risk maps to synaptic biology but higher functional resolution is needed to understand the disease mechanisms.
3. Lymphoblast cell lines make for a useful experimental tool to study adverse molecular medication effects associated with clozapine.
4. My research on clozapine yielded more questions than answers – there is still a lot we do not know about the molecular function of the antipsychotic drug.
5. DNA methylation algorithms are promising new genomic tools but require careful implementation and calculations to maximize their potential.
6. DNA methylation age is affected in schizophrenia with age- and sex-specific effects.
7. Blood-based DNA methylation algorithms hold promise as clinical tools for schizophrenia, but more research and funding are needed.
8. Open and collaborative science is the path forward to decipher the biology of schizophrenia, and in scientific research as a whole.

Samenvatting

Een op de 200 mensen krijgt ergens in hun leven een diagnose van schizofrenie en 1.5 miljoen mensen hebben wereldwijd dit jaar een schizofrenie diagnose gekregen. Ondanks de enorme impact van de ziekte op het lichaam en leven van mensen, hun dierbaren, en onze samenleving, is onze kennis over de biologie van schizofrenie heel beperkt. Daarbovenop is er momenteel ook nog geen geneesmiddel beschikbaar. Nieuwe ontwikkeling in genomische technologieën en in methodologie in de kwantitatieve genetica hebben het afgelopen decennium een belangrijke rol gespeeld in nieuwe kennis die is vergaard over de biologie van complexe ziektes, zoals schizofrenie dat is. Grootschalige genetische studies hebben bijvoorbeeld specifieke regio's in ons DNA geidentificeerd, die een mens een verhoogde kans geven voor het ontwikkelen van de ziekte. Dit is een uiterst belangrijke mijlpaal in psychiatrisch genetisch onderzoek, omdat het ons voor eerst een kijkje geeft in mogelijke genetische oorzaken die ten grondslag liggen aan schizofrenie. Vervolgonderzoek is nu nodig dat de functionele mechanismen verder ontrafeld en daarmee de moleculaire en cellulaire processes kan verhelderen die een rol spelen bij de ziekte. Tegelijkertijd is het net zo belangrijk om de moleculaire gevolgen, die een ernstige impact maken op het lichaam en het leven van patiënten, ook beter in kaart te brengen.

Deze dissertatie is een verzameling van mijn onderzoek naar de biologie van schizofrenie. Het representeert de technologische ontwikkeling en diversiteit in genomisch onderzoek die tot onze beschikking staat om het ziektebeeld van schizofrenie en diens moleculaire gevolgen beter in kaart te brengen. Met behulp van kweekmodellen van neuronale en bloedcellen, grootschalige case-control cohorten, recent ontwikkelde DNA methylatie algoritmes, en hergebruik van al gepubliceerde genomische datasets, heb ik een diepgaand onderzoek naar de biologie van schizofrenie uitgevoerd. Mijn doel was om voort te borduren op de uitkomsten van grootschalige genetische studies en door gebruik van de nieuwste methodiek en integratieve genomische analyses de volgende stap te nemen in het beter begrijpen van de biologie van schizofrenie. Met mijn dissertatie presenteer ik acht conclusies die ik van belang acht voor onze kennis over de ziekte of voor het ondernemen van genetisch onderzoek naar psychiatrische ziekten in het algemeen.

Het verhelderen van de biologie van schizofrenie is een essentiële stap naar het ontwikkelen van nieuwe behandelingen en het humaniseren van deze ernstige ziekte die nog steeds gestigmatiseerd wordt in onze samenleving. Mijn bevindingen bieden belangrijke bouwstukken voor vervolgonderzoek om de moleculaire oorzaken en gevolgen van de ziekte beter te begrijpen. Alhoewel er nog veel uitdagingen te overwinnen zijn, laat mijn onderzoek zien dat de wegen naar oplossingen en daarmee ook onze hoop op een beter leven voor patiënten nog nooit zo rijk zijn geweest.

Mijn acht belangrijkste conclusies:

1. In functioneel onderzoek naar schizofrenie zullen wij diens polygene architecture moeten omarmen.
2. Het genetisch risico van schizofrenie is geassocieerd met synaptische genen en pathways maar een hogere functionele resolutie is nodig om het ziektemechanisme beter te kunnen begrijpen .
3. Kweekmodellen van lymfoblastoide cellijnen zijn een bruikbaar experimenteel handvat om de moleculaire bijwerkingen van clozapine te onderzoeken.
4. Mijn onderzoek naar clozapine heeft meer vragen dan antwoorden boven water gebracht - we weten nog steeds weinig over de precieze werking van deze antipsychotische medicatie.
5. DNA methylatie algoritmes zijn nieuwe veelbelovende genomische methoden die wel zorgvuldige berekeningen nodig hebben voor een betrouwbare implementatie.
6. DNA methylatie leeftijd is aangetast in schizofrenie op een manier die verschillen laat zien per leeftijd en geslacht.
7. DNA methylatie algoritmes die gebruikmaken van meting uit bloed hebben de potentie een klinisch hulpmiddel te worden voor de behandeling van schizofrenie, maar hiervoor zal eerst meer onderzoek en financiering nodig zijn.
8. Een wetenschap waarin transparantie, toegankelijkheid en samenwerking centraal staan, is het pad voorwaarts om schizofrenie op te lossen, maar ook voor wetenschappelijk onderzoek in het algemeen.

Propositions

1. GWAS SNP-heritability is a population parameter that can be experimentally studied in a laboratory dish.
2. Blood-based DNA methylation data is an undervalued resource in psychiatric genomic research.
3. Replication of results should be the aim in genomic research and not optimization of data processing pipelines, as technical confounding is inevitable in any -omic analyses.
4. Schizophrenia may indeed be a syndrome of aging with biological age being altered both intrinsically as part of its etiology as well as by consequences of the illness that compound over time.
5. To improve the therapeutic impact of the antipsychotic clozapine, drug modification efforts should focus on reducing its molecular adverse effect on cholesterol-related genes and pathways in blood cells.
6. Most GWAS findings reflect variants and genes with no direct relevance for the underlying biology of the phenotype studied.
7. Genetic causality is rarely deterministic for complex traits.
8. The illusion of meritocracy in academia hinders scientific progress that can be achieved when researchers operate as a unified collective.
9. Science that does not foster diversity, equity, and inclusion at its core, is not science but is oppression.
10. Data that is collected by government research funding should be made open access and available to all stakeholders immediately.
11. This thesis is not heritable but GWAS will likely find a genetic association.

About the author

Anil Pravin Surendredath Ori was born in Paramaribo, Suriname, on January 4, 1987. Together with his parents and older brother, he immigrated to The Hague, the Netherlands, in 1995. After completing his gymnasium high school education at Dalton Den Haag, in The Hague, he embarked on his undergraduate university studies at the Universiteit van Amsterdam (UvA). Here, he completed an interdisciplinary course curriculum in the Life Sciences, including an academic exchange program and visit to the University of Toronto, in Canada, during the winter of 2007, before graduating at the UvA in 2009. He started his master's studies at Universiteit Utrecht, later that year. There he completed the Biology of Disease research track within the Biomedical Sciences program with a major in Human Genetics and Neuroscience and a minor in Business and Entrepreneurship. During his studies in Utrecht, he conducted full-time research at the Human Genetics department. In collaboration with Prof. Dr. Roel A. Ophoff, he studied gene expression changes in human postmortem brain samples of patients diagnosed with psychiatric illnesses. After successful completion of his master studies, he immigrated to Los Angeles, California, to continue his research in psychiatric genetics. At the Center for Neurobehavioral Genetics at the University of California, Los Angeles (UCLA), he worked as a research scientist and data analyst in the laboratory of Prof. Dr. Ophoff and conducted his doctoral research studying the genomic causes and consequences of schizophrenia.

During his PhD work, he was awarded awards and prizes, including a Finalist Best Poster Award at the 2014 World Congress Psychiatric Genetics annual meeting in Copenhagen, Denmark, a Reviewer's Choice Abstract Award at the 2015 American Society of Human Genetics (ASHG) annual meeting in Baltimore, USA, an Early Career Investigator Travel Award and Best Poster Award at the 2016 WCPG annual meeting in Jerusalem, Israel, and a semi-finalist Charles J. Epstein Trainee Award for Excellence in Human Genetics Research at the 2016 ASHG annual meeting in Vancouver, Canada. He was also a selected attendee at the Leena Peltonen School of Human Genomics summer school of 2018, in Les Diablerets, Switzerland. In addition to publishing his research in scientific journals, he has presented his work at both national and international workshops and conferences and supervised undergraduate and graduate students on their research and thesis projects.

After his time at UCLA, he relocated back to the Netherlands where he continued his research on the genomics of psychiatric illnesses at the Department of Psychiatry and the Department of Genetics at the University Medical Center Groningen, in Groningen. He finished writing his doctoral thesis in his own time which he will soon defend at Erasmus University Rotterdam where he is enrolled as an external Ph.D. candidate at the Department of Psychiatry of the Erasmus Medical Center.

PhD Portfolio

Name:	Anil P.S. Ori
Department at Erasmus Medical Centre:	Psychiatry
Research School:	Erasmus Medical Center Graduate School
PhD period:	October 2013 – July 2021
Promotor:	Prof. Dr. Roel A. Ophoff
Co-promotor:	Prof. Dr. Steven Kushner

1. PhD Training	year	workloaf (ECTS)
Course work		
Computational Genetics UCLA CS/HG 124/224	2013	5
Introduction to Computer Science and Programming (EdX-MITx)	2013	2
Gene Network Analysis Short Course (UCLA)	2013	1.5
Statistical Methods in Computational Biology UCLA M271	2014	5
Cell, Development, and Molecular Neurobiology UCLA M201	2015	5
Current Research Topics in Neurobehavioral Genetics UCLA NS215	2016	1
Workshops and Summer Schools		
UCLA PhD Career Conference/Workshops	2014/17	1
UCLA ICNN RNA-sequencing Workshop	2014	0.5
UCLA Computational Genomics Summer School	2016	3
Leena Peltonen School of Human Genomics	2018	2
Other training		
Psychiatric Genetics Journal Club (biweekly)	2013-2019	5
Project Brainstorm UCLA	2016	0.25

2. Conference participation		
Poster presentations		
World Congress of Psychiatric Genetics, Boston, MA, USA	2013	1
American Society of Human Genetics, Boston, MA, USA	2013	1
Society for Neuroscience, San Diego, CA, USA	2013	1
The R User Conference, Los Angeles, CA, USA	2014	0.5
World Congress of Psychiatric Genetics, Copenhagen, Denmark	2014	1
World Congress of Psychiatric Genetics, Toronto, Canada	2015	1
American Society of Human Genetics, Baltimore, MD, USA	2015	1
World Congress of Psychiatric Genetics, Jerusalem, Israel	2016	1
American Society of Human Genetics, Vancouver, BC, Canada	2016	1
World Congress of Psychiatric Genetics, Florida, FL, USA	2017	1
World Congress of Psychiatric Genetics, Glasgow, Scotland	2018	1

Platform presentations

American Society of Human Genetics, Vancouver, BC, Canada	2018	0.25
World Congress of Psychiatric Genetics, Los Angeles, CA, USA	2019	0.25

3. Seminar and invited presentations

UCLA Neurobehavioral Genetics Retreat	2016	0.25
UCLA NIDP Graduate Recruitment	2016	0.25
SGDP Statistical Genetics Department, King's College London, UK	2016	0.25
UCLA Bruin Talk Seminar series	2017	0.25
UCLA Medical and Population Genomics Seminar Series	2017	0.25
UCLA Medical and Population Genomics Seminar Series	2018	0.25
UCLA Medical and Population Genomics Seminar Series	2018	0.25

4. Teaching and mentoring activities

Supervision Master's theses

Student: Remco Molenhuis Program: Neuroscience and Cognition, Utrecht University Project: Neurodevelopment and microRNA-137in stem cell models	2013	5
Student: Merel Bot Program: Neuroscience and Cognition, Utrecht University Project: Towards identifying microRNA-137 targets across neurodevelopment	2014/15	5
Student: Anouk Verboven Program: Drug Innovation, Utrecht University Project: Changes in mtDNA content in bipolar patients, their siblings and its relation to clinical factors	2016	5

Supervision Undergraduate Theses

Student: Vidhi Rao Program: Biochemistry, UCLA Project: Detection of ALS-associated C9ORF72 Repeat Expansion by Southern Blot Using the DIG-Labeling System	2013/14	5
Student: Daniel Bandary Program: Microbiology, Immunology, & Molecular Genetics, UCLA Project: The Neurexin-1 Gene: CNV MLPA validation	2014	5

5. Professional memberships and activities

Member of the American Society of Human Genetics	2013-2019	-
Member of the International Society of Psychiatry Genetics	2013-present	-
Member of the Federation of European Neuroscience Societies	2013-2019	-
Member of the Society for Neuroscience	2013	-
Ad hoc reviewer	2013-present	3
Rapporteur World Congress Psychiatric Genetics meeting	2016	0.25

6. Communities and Service

Second Round Judge, ASHG's Annual DNA Day Assay Contest	2014-2019	0.5
Volunteer UCLA DNA Day, Los Angeles, CA, USA	2016	0.25
Volunteer UCLA Brain Awareness Week, Los Angeles, CA, USA	2016	0.5
Member of Project Synapse, UCLA Brain Research Institute	2016-2019	2
Graduate Student Mentor UCLA HIV Counseling and Testing Coalition	2017-2018	2
Volunteer Los Angeles/Irvine Brain Bee competition, CA, USA	2017-2019	0.5

Total ECTS = 79.5

1 ECTS (European Credit Transfer System) is equal to a workload of 28 hours

List of publications

A. Thesis research chapters

1. **Ori APS**, Bot MHM, Molenhuis RT, Olde Loohuis LM, Ophoff RA. A longitudinal model of human neuronal differentiation for functional investigation of schizophrenia polygenic risk. Biological Psychiatry. 2019 Apr 1;85(7):544-553. [PMID: 30340753]

2. **Ori APS***, de With SAJ*, Pulit SJ, Wang T, Strengman E, Glennon JC, Buitelaar JK, Viana J, Staal WG, de Jong S, Ophoff RA. Integrative genomic strategies applied to a lymphoblast cell line model reveal specific transcriptomic signatures associated with clozapine response. Submitted for publication. BioRxiv 2020 Sep:
https://doi.org/10.1101/2020.09.22.308262.
**Equal contribution*

3. **Ori APS**, Olde Loohuis LM, Guintivano J, Hannon E, Dempster E, St. Clair D, Bass NJ, McQuillin A, Mill J, Sullivan P, Kahn RS, Horvath S, Ophoff RA. Epigenetic age is accelerated in schizophrenia with age- and sex-specific effects and associated with polygenic disease risk. Submitted for publication. BioRxiv 2021 Feb:
https://doi.org/10.1101/727859.

4. **Ori APS**, Lu AT, Horvath S, Ophoff RA. A systematic evaluation of 41 DNA methylation predictors across 101 data preprocessing and normalization strategies highlights considerable variation in algorithm performance. Submitted for publication. BioRxiv 2021 Oct:
https://doi.org/10.1101/2021.09.29.462387.

B. Highlighted in thesis discussion chapter

5. Teeuw J, **Ori APS**, Brouwer RM, de Zwarte SMC, Schnack HG, Hulshoff Pol HE, Ophoff RA. Accelerated aging in the brain, epigenetic aging in blood, and polygenic risk for schizophrenia. Schizophrenia Research. 2021 May 231:189-197. [PMID: 33882370]

6. Krebs CE, **Ori APS**,Vreeker A, Wu T, Cantor RM, Boks MP, Kahn RS, Olde Loohuis LM, Ophoff RA. Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect. Psychological Medicine, 2020 Nov, pp. 2575 - 2586. [PMID: 31589133]

7. Jia T, Chu C, ... **Ori APS**, ..., Thompson PM, Schumann G, Desrivières S. Epigenome-wide meta-analysis of blood DNA methylation and its association with subcortical volumes: findings from the ENIGMA Epigenetics Working Group. Molecular Psychiatry, 2019 Dec 6. [PMID: 31811260]

8. Kowalec K, Hannon E, Mansell G, Burrage J, **Ori APS**, Ophoff RA, Mill J, Sullivan PF. Methylation age acceleration does not predict mortality in schizophrenia. Translational Psychiatry. 2019 Jun 4;9(1):157. [PMID: 31164630]

9. Stahl EA, Breen G, ..., **Ori APS**, ..., Ophoff RA, Scott LJ, Andreassen OA, Kelsoe J & Sklar P. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nature Genetics. 2019 May;51(5):793-803. [PMID: 31043756]
10. 10. Ruderfer DM, Ripke S, ..., **Ori APS**, ..., Ophoff RA, Wray NR, Sklar P, Kendler KS. Genomic dissection of bipolar disorder and schizophrenia including 28 subphenotypes. Cell. 2018 173:1705–1715. [PMID: 29906448]

11. Olde Loohuis LM, Vorstman JAS, **Ori APS**, Staats KA, Wang T, Richards AL, Leonenko G, Walters JT, DeYoung J, GROUP consortium, Cantor RM & Ophoff RA. Genome-wide burden of deleterious coding variants increased in schizophrenia. Nature Communications. 2015 6:7501. [PMID: 26158538]

C. Not part of thesis

12. Ballering AV, **Ori APS**, Rosmalen JGM. The association of sex, age and FKBP5 genotype with common somatic symptoms: A replication study in the lifelines cohort study. Journal of Psychosomatic Research 2021 Aug; 147:110510. [PMID: 34034139]

13. Mc Intyre K*, Lanting P*, Deelen P*, Wiersma H*, Vonk JM*, **Ori APS***, Jankipersadsing SA*, ... , Jackie Dekens, Jochen O. Mierau, H. Marike Boezen, Lude Franke. The Lifelines COVID-19 Cohort: a questionnaire-based study to investigate COVID-19 infection and its health and societal impacts in a Dutch population-based cohort. BMJ Open 2021 Mar 17;11(3):e044474. [PMID: 33737436]
**Contributed equally*

14. Olde Loohuis LM, Mennigen E, **Ori APS**, Perkins D, ..., Bearden CE & Ophoff RA. Genetic and clinical analyses of psychosis spectrum symptoms in a large multiethnic youth cohort reveal significant link with ADHD. Translational Psychiatry 2021 Jan 28;11(1):80. [PMID: 33510130]

15. van Blokland I*, Lanting P*, **Ori APS***, Vonk JM*, Warmerdam RC*, ..., Deelen P, Boezen HM, Franke LH, Lifelines COVID-19 cohort study, The COVID-19 Host Genetics Initiative. Using symptom-based case predictions to identify host genetic factors that contribute to COVID-19 susceptibility. Submitted for publication. MedRxiv 2020 Aug:
https://doi.org/10.1101/2020.08.21.20177246

16. Sul JH, Service S, ..., **Ori APS**, ..., Carrie Bearden, Chiara Sabatti, Nelson Freimer. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. Translational Psychiatry 2020 Feb 24;10(1):74. [PMID: 32094344]

17. Lee. Y, Sun D, **Ori APS**, Lu AT, ..., Aviv A, Jugessur A, Horvath S. Epigenome-wide association study of leukocyte telomere length. Aging 2019 Aug 26;11(16):5876-5894. [PMID: 31461406]

18. Olde Loohuis LM, Mangul S, **Ori APS**, Jospin Guillaume, Koslicki D, Wu T, Boks MP, Lomen-Hoerth C, Wiedau-Pazos M, Cantor RM, de Vos WM, Kahn RS, Eisen JA, Eskin E, Ophoff RA. Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. Translational Psychiatry. 2018 May 10;8(1):96. [PMID: 29743478]

19. Ciobanu LG, **Ori APS**, Pagliaroli L, Polimanti R, Spindola LM, Vincent JB, Cormack FK. Summaries of Plenary and Selected Symposia Sessions at the XXIV World Congress of Psychiatric Genetics, Jerusalem, Israel, October 30 – November 3, 2016. Psychiatric Genetics. 2017 Apr 27(2):41-53. [PMID: 28212207]

20. Luykx JJ, Bakker SC, Visser WF, Verhoeven-Duif N, Buizer-Voskamp JE, den Heijer JM, Boks MP, Sul JH, Eskin E, **Ori APS**, Cantor RM, Vorstman J, Strengman E, DeYoung J, Kappen TH, Pariama E, van Dongen EP, Borgdorff P, Bruins P, de Koning TJ, Kahn RS, Ophoff RA. Genome-wide association study of NMDA receptor coagonists in human cerebrospinal fluid and plasma. Molecular Psychiatry. 2015 Dec 20(12):1557-64. [PMID: 25666758]

21. de Jong S, Boks MP, Fuller TF, Strengman E, Janson E, de Kovel CG, **Ori APS**, Vi N, Mulder F, Blom JD, Glenthøj B, Schubart CD, Cahn W, Kahn RS, Horvath S, Ophoff RA. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. PLOS ONE. 2012;7(6):e39498. [PMID: 22761806]

Acknowledgements

I owe a great debt of gratitude to all of you who supported me over the past years. Without you I would not have been able to accomplish the research described in this thesis. I would therefore like to conclude this book by expressing my sincere gratitude to and appreciation for those that have been part of this journey.

My beloved family and chosen family. I am because we are. My heart warms for you and because of you. My gratitude is beyond words. I am so thankful for and appreciative of your love and support. Thank you for everything. I dedicate this book to all of you.

My dear mentors, both those within as those outside of academia. Thank you for sharing your knowledge and experience and thank you for your patience. Your guidance and constructive feedback have shaped me to be the person and scientist I am today. I am so grateful for your support. This work is a tribute to you as well.

My wonderful colleagues. I can't express how much I enjoyed working with you and learning from you. Thank you for the great collaborations, all the brainstorm sessions, the coffee and lunch breaks, the drinks and dinners, and all the conference adventures. A special shout out goes to my colleagues on Twitter who provided opportunities to learn and have a laugh online as well. I feel so rich to have been able to work with you and hope our paths will cross again in the future.

My beautiful friends. Spread across the world, sometimes for longer periods far apart, but when we see and speak to each other, we are connected in friendship and great company. I am here because of your love and support. Thank you for being so amazing, so kind, so caring, and so loving.

The heroes who are often forgotten. To my colleagues who work in the support teams and who facilitate and arrange much of the needed infrastructure and paperwork to do research. To my colleagues who work in cafes and restaurant in university and research buildings. To my colleagues who work in cleaning and keep the buildings and offices tidy and clean on a daily. To all study participants who often offer their precious time and (biological) data so kindly and generously to advance science and help others. To the many students who take on so much work without getting paid. To the many colleagues who work so much without getting paid. To my colleagues who make their softwares, their analytical code, their collected data publicly accessible so others can use and build further up on their work. To my colleagues who upload their work on preprint servers so scientific knowledge is immediately accessible to everyone with an internet connection. I am so thankful to and grateful for all of you. My research would not have been possible without your help and labor. They say that science stands on the shoulders of giants, that we build on the work of previous scientist. In practice, however, science stands on your shoulders. The true heroes deserving of a Nobel Prize.

Dankwoord
Mijn dankbaarheid is groot aan eenieder die mij gesteund heeft over de afgelopen jaren. Zonder jullie was dit proefschrift niet tot stand gekomen. Ik wil dit boek daarom afsluiten door mijn dank aan en waardering voor jullie te benadrukken.

Mijn lieve familie en gekozen familie. Ik ben omdat wij zijn. Mijn hart is groot voor en door jullie. Mijn dankbaarheid oneindig. Jullie zijn mij zo dierbaar. Door jullie onvoorwaardelijk steun voelt het leven als de wind mee hebben. Dankjewel voor alles en dit boek draag ik op aan jullie.

Mijn beste mentoren, zowel die binnen als buiten de academie. Bedankt voor het delen van jullie kennis en ervaring en voor jullie geduld ook. Jullie mentorschap en constructieve feedback heeft mij gevormd tot de persoon en onderzoeker die ik nu ben. Ik ben ontzettend dankbaar dat ik op jullie kan rekenen. Mijn werk is ook jullie werk.

Mijn fantastische collega's. Wat is het enorm fijn om met jullie te werken en om van jullie te leren. Dank jullie wel voor alle samenwerkingen, alle brainstormsessies, alle koffie en lunchpauzes, alle borrels, en alle congresavonturen. Een speciale shout out ook naar mijn collega's op Twitter die ook online voor veel leerzame momenten en plezier hebben gezorgd. Ik voel me erg rijk dat ik met jullie heb mogen werken en hoop dat onze wegen elkaar weer zullen kruisen.

Mijn dierbare vrienden. Verspreid over de wereld, in tijden soms ver van elkaar, maar als we elkaar zien of spreken, verbonden in goed gezelschap en vriendschap. Ik zou hier niet zijn zonder jullie steun en liefde. Dank jullie wel dat jullie zo attent zijn, zo zorgzaam, zo gezellig, en zo liefdevol.

De vaak vergeten helden. De collega's die in de administratie werken en al het papierwerk in goede banen leiden. De collega's die de eetcafés en restaurants op universiteiten en onderzoeksinstituten draaiende houden. De collega's die de gebouwen en kantoren schoonmaken. Alle studiedeelnemers die hun waardevolle tijd en (biologische) data zo onbaatzuchtig doneren aan de wetenschap om anderen te helpen. De vele studenten die zo hard werken en onbetaald werk uitvoeren. De vele collega's die onbetaald werk uitvoeren en overuren maken. De collega's die hun softwares, hun analyse code, en hun verzamelde data openbaar maken zodat anderen daarop voort kunnen bouwen. Collega's die hun werk op preprint servers plaatsen zodat wetenschappelijke kennis voor iedereen met een internetverbinding direct toegankelijk is. Mijn dankbaarheid is groot aan jullie allen. Ik heb mijn werk kunnen uitvoeren dankzij jullie hulp en arbeid. Er wordt weleens gezegd dat de wetenschap staat op de schouders van reuzen, op het werk van wetenschappers die er voor ons waren. Maar in werkelijkheid staat de wetenschap op de schouders van jullie. De ware helden die een Nobelprijs verdienen.

This dissertation embodies the genomic opportunities and diverse research strategies that lay at our disposal to improve our understanding of schizophrenia, a major psychiatric disorder that affects millions of people worldwide. Through my research, I aimed to go beyond the findings of large-scale genetic studies and conducted research that uses state-of-the-art methodology and integrative genomic data analyses to shed light on the biology of schizophrenia.

On a personal level, my aim was to learn about the illness and deepen my understanding of its phenotypic and biological characteristics and complexity. Having loved ones who have been diagnosed with the illness, I have seen the impact on and suffering of those affected up close. This research has in part been a journey of understanding pain and trauma in my life and expanding the depths of my support and love for those who suffer and have suffered.

The photo on the front shows the arms of my mother and myself in union. With an illness that is so stigmatized, I felt it was important to communicate warmth and hope. As we wait on a cure to be discovered, we rely on what we know are protective factors; family and community support.