
Search for genetic variants associated with schizophrenia using next-generation sequencing

MAXIMILIAN HASTREITER

Lehrstuhl für Genomorientierte Bioinformatik

TECHNISCHE UNIVERSITÄT MÜNCHEN

2019



Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

**Search for genetic variants associated with schizophrenia using
next-generation sequencing**

Maximilian Hastreiter

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Frank Johannes

Prüfer der Dissertation:

1. Prof. Dr. Hans-Werner Mewes
2. Prof. Dr. Dmitrij Frischmann

Die Dissertation wurde am 03.09.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 18.12.2019 angenommen.

Für Katharina und Tamara

DANKSAGUNG

An erster Stelle gilt mein Dank vor allem meinem Doktorvater Werner Mewes. Ich möchte mich hierbei insbesondere für die Ermöglichung meiner Dissertation sowie für die durchgehende Unterstützung meiner wissenschaftlichen Laufbahn herzlich bedanken.

Desweiteren gilt mein besonderer Dank Robert Küffner für die langjährige Betreuung und Unterstützung aus dem In- und Ausland, mit deren Hilfe viele Projekte und Umsetzungen erst ermöglicht wurden.

Mein Dank gilt auch meinen Kollegen und Kooperationspartnern aus anderen Instituten die mich unterstützt und in ihre Projekte involviert haben. Insbesondere möchte ich mich dabei bei Peter Huypens (IEG) für die zahlreichen Diskussionen und Projekte bedanken, ohne die meine Zeit sicherlich ein Stück weniger interessant gewesen wäre. Ebenso gilt mein ausdrücklicher Dank Uli Güldener (TUM WZW) der meinen thematischen Ausflug in das Reich der Pilze begleitet und unterstützt hat.

Ich möchte mich ausserdem bei Gabi Kastenmüller sowie den gesamten Kollegen am IBIS für das, in meinen Augen, aussergewöhnlich kollegiale und freundliche Arbeitsumfeld bedanken.

Zu guter Letzt möchte ich mich bei meinen Doktoranden- und Zimmerkollegen Jonathan Hoser, Tim Jeske, Sven Duscha und Patrick Dreher bedanken, mit denen ich einige unvergessliche Momente erleben durfte.

ABSTRACT

Schizophrenia is one of the most severe mental diseases. Prevalence is estimated with about five per thousand in the population [1]. The disease is often associated with severe symptoms such as delusions, disorganised speech and thought, and social withdrawal. Despite the efforts of numerous studies to investigate the causes of the disease, the underlying mechanisms are still unknown. Still it has been shown that genetics plays a central role in etiology, with heritability estimates of about 80%. In this thesis, the analysis of whole-exome sequencing data from 37 trios and 210 other patients will be used to find evidence for known and novel candidate genes that might help to explain this complex phenotype.

I used a unique combination of bioinformatic methods, including the study of *de novo*, loss-of-function and compound heterozygous variants to carry out this work. In addition, new approaches such as the Edgetics method were applied to find a comprehensive, high quality set of candidate genes. This strategy made it possible to jointly analyse different aspects of the genetic spectrum and obtain a picture of the mutational load.

In the ensuing discussion, I show that the detected genetic variations particularly disrupt parts of the neuronal cytoskeleton which is essential for a variety of neuronal processes. Foremost this is shown by damaging genetic mutations found in genes related to microtubules and the actin cytoskeleton. Functionally, the variants can be assigned to certain neuronal processes such as stability control, transport or cell migration. Taken together, these results indicate that dysregulations and disruptions within these central mechanisms might play an important role in the pathogenesis of schizophrenia.

In the last part of my work, I discuss important technical aspects that are essential for the evaluation of the obtained results. This includes basic strategies such as the removal of PCR duplicates and the determination of haplotypes.

In summary, my thesis provides a detailed genetic analysis of a large cohort of schizophrenia patients. With the implementation of an unique analysis strategy that combines various computational methods, I was able to detect an accumulation of deleterious events, particularly affecting the neuronal cytoskeleton. Finally, I conduct a critical discussion, that allows to assess the presented results on a biological and technical level.

ZUSAMMENFASSUNG

Schizophrenie ist eine der schwersten psychischen Erkrankungen. Die Prävalenz wird auf etwa fünf Betroffene pro 1000 Personen geschätzt [1]. Die Erkrankung ist häufig mit schweren Symptomen wie Wahnvorstellungen, unorganisiertem Sprechen und Denken sowie sozialem Rückzug verbunden. Trotz der Bemühungen zahlreicher Studien, die die Ursachen der Krankheit untersuchten, sind die zugrunde liegenden Mechanismen noch immer unklar. Es ist dennoch akzeptiert, dass die Genetik eine zentrale Rolle in der Ätiologie der Schizophrenie spielt, da Schätzungen der Erbllichkeit von etwa 80% ausgehen. In dieser Arbeit soll nun durch die Analyse von Whole-Exome Sequenzierungsdaten von 37 Trios sowie 210 weiteren Patienten Hinweise für bekannte und neuartige Kandidatengene gefunden werden, die helfen könnten, diesen komplexen Phänotyp zu erklären.

Zur Durchführung dieser Arbeit wurde eine einzigartige Kombination von bioinformatischen Methoden, einschließlich der Untersuchung von *de novo*, Loss-of-Function und compound heterozygoten Varianten verwendet. Zusätzlich wurden neue Ansätze wie die Edgetics-Methode verwendet um einen umfassenden, qualitativ hochwertigen Satz von Kandidatengenen zu finden. Diese Strategie ermöglichte es, verschiedene Aspekte des genetischen Spektrums gemeinsam zu analysieren und ein Bild der gesamten Mutationslast zu erhalten.

In der anschließenden Diskussion zeige ich, dass die gefundenen genetischen Variationen Teile des neuronalen Zytoskeletts zerstören, das für eine Vielzahl von neuronalen Prozessen unerlässlich ist. Hierbei waren vor allem Mikrotubuli und das Aktin-Zytoskelett betroffen. Funktionell können die Varianten bestimmten neuronalen Prozessen wie der Stabilitätskontrolle, Transport oder Zell-Migration zugeordnet werden. Zusammenfassend, deuten diese Ergebnisse darauf hin, dass Dysregulationen und Störungen innerhalb dieser zentralen Mechanismen eine wichtige Rolle bei der Pathogenese der Schizophrenie spielen.

Im letzten Teil meiner Arbeit diskutiere ich weitere wichtige technische Aspekte, die für die Bewertung der erzielten Ergebnisse unerlässlich sind. Dazu gehören grundlegende Strategien wie die Entfernung von PCR-Duplikaten und die Bestimmung von Haplotypen.

Zusammenfassend liefert die vorliegende Arbeit eine detaillierte genetische Analyse und Diskussion einer Kohorte von Schizophreniepatienten. Mit der Implementierung einer einzigartigen Analysestrategie, die verschiedene bioinformatische Methoden kombiniert, konnte ich eine Anreicherung von schädlichen Mutationen im neuronale Zytoskelett feststellen. Abschließend führe ich eine kritische Diskussion, die es ermöglicht, die präsentierten Ergebnisse auf einer

biologischen und technischen Ebene zu bewerten.

SCIENTIFIC CONTRIBUTIONS

The following list specifies all scientific contributions including peer-reviewed publications and conference contributions.

- Hastreiter, M., Küffner, R., Giegling, I., Hartmann, A.M., Konte, B., Jeske, T., Hoser, J., Spitzmüller, A., Mewes, H.-W., Rujescu, D., (*submitted*). **Comprehensive whole-exome sequencing analysis of sporadic schizophrenia indicates disruptions in the neuronal cytoskeleton**
- Bosch, J., Czedik-Eysenberg, A., Hastreiter, M., Khan, M., Güldener, U., Djamei, A., (*Molecular Plant-Microbe Interactions, accepted*). **Two is better than one: Using a hybrid fungus to understand host specificity on a model grass species**
- Jeske, T., Huypens, P., Stirn, L., Höckele, S., Wurmser, C., Böhm, A., Weigert, C., Staiger, H., Klein, C., Beckers, J., Hastreiter, M., 2019. **DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences.** Bioinformatics. doi:10.1093/bioinformatics/btz495
- Hastreiter, M., Jeske, T., Hoser, J., Kluge, M., Ahomaa, K., Friedl, M.-S., Kopetzky, S.J., Quell, J.-D., Mewes, H.-W., Küffner, R., 2017. **KNIME4NGS: a comprehensive toolbox for Next Generation Sequencing analysis.** Bioinformatics. doi:10.1093/bioinformatics/btx003
- Hastreiter, M., Jeske, T., Rujescu, D., Küffner, R., Mewes, H.-W., **Deciphering the genetic background of schizophrenia based on WES data analysis.** Poster session at the 16th Eibsee Meeting. October 2016, Germany.
- Albanese, M., Tagawa, T., Bouvet, M., Maliqi, L., Lutter, D., Hoser, J., Hastreiter, M., Hayes, M., Sugden, B., Martin, L., Moosmann, A., Hammerschmidt, W., 2016. **Epstein-Barr virus microRNAs reduce immune surveillance by virus-specific CD8+ T cells.** Proc. Natl. Acad. Sci. U.S.A. 113, E6467-E6475. doi:10.1073/pnas.1605884113

-
- Tagawa, T., Albanese, M., Bouvet, M., Moosmann, A., Mautner, J., Heissmeyer, V., Zielinski, C., Lutter, D., Hoser, J., Hastreiter, M., Hayes, M., Sugden, B., Hammerschmidt, W., 2016. **Epstein-Barr viral miRNAs inhibit antiviral CD4+ T cell responses targeting IL-12 and peptide processing.** J. Exp. Med. 213, 2065-2080. doi:10.1084/jem.20160248
 - Hoser, J., Hastreiter, M., Jeske, T., Kluge, M., Friedl, M.-S., Kopetzky, S.J., Quell, J.-D., Küffner, R., Mewes, H.-W., **KNIME for NGS - The Next Generation.** Poster session at the KNIME UGM. February 2015, Berlin, Germany.
 - Hastreiter, M., Rujescu, D., Giegling, I., Hoser, J., Küffner, R., Mewes, H.-W., **Identification of Loss-of-Function variants in exome data.** Poster session at the Gordon Conference: Human Single Nucleotide Polymorphisms & Diseases. August 2014, Easton, USA.

TABLE OF CONTENTS

	Page
1 Introduction	1
1.1 Motivation	1
1.2 Mental diseases, a global burden	2
1.3 Schizophrenia	4
1.3.1 The definition of the term "schizophrenia"	4
1.3.2 Classification, symptoms and subtypes	5
1.3.3 Pathophysiology	8
1.3.4 Etiology	10
1.3.5 Genetics of schizophrenia	11
2 Materials and Methods	15
2.1 Datasets and resources	15
2.1.1 The m4-Schizophrenia sequencing project	15
2.1.2 External reference resources	16
2.2 Variant calling and phasing pipeline	16
2.2.1 Data preprocessing	16
2.2.2 Read mapping and alignment processing	18
2.2.3 Variant calling and filtering	19
2.2.4 Haplotype phasing	22
2.3 Variant annotation and analysis workflow	23
2.3.1 The Combined Annotation Dependent Depletion (CADD) score	23
2.3.2 <i>De novo</i> variant calling	24
2.3.3 <i>De novo</i> gene knockouts	26
2.3.4 Genes enriched for deleterious variants	26
2.3.5 Edgetics - a link between genotype and phenotype	28
2.3.6 Gene set over-representation enrichment analysis	29
2.3.7 Protein-protein interaction networks	30
2.4 Structural variants and known candidates	30
2.4.1 Detection of long insertions and deletions using Pindel	30

TABLE OF CONTENTS

2.4.2	ClinVar, an archive for relationships among sequence variations and human phenotypes	31
2.4.3	The International Mouse Phenotyping Consortium	32
2.5	The Konstanz information miner	32
2.5.1	KNIME4NGS	33
3	Results	35
3.1	Data quality, filtering and mapping	35
3.1.1	Read filtering	35
3.1.2	Read mapping and target region coverage	36
3.1.3	Alignment processing, quality filtering and variant calling	36
3.2	Analysis of genetic variation types	36
3.2.1	<i>De novo</i> variants	40
3.2.2	Compound heterozygous variants	41
3.2.3	Loss-of-Function variants	43
3.2.4	Genetic Burden	44
3.2.5	Edgetics	45
3.3	Candidate genes and integrative analysis	48
3.3.1	Protein-protein interaction network	48
3.3.2	Enrichment analysis	49
3.3.3	Genes with schizophrenia-related mouse phenotypes	52
3.4	Analysis of structural variants and association with known candidates	53
3.4.1	Structural variants on chromosome 22	53
3.4.2	ClinVar	54
4	Discussion	59
4.1	Comprehensive whole-exome sequencing analysis of sporadic schizophrenia	59
4.1.1	Rate of <i>de novo</i> variants supports important role in schizophrenia	59
4.1.2	Rare deleterious compound heterozygous variants affect a broad range of genes	60
4.1.3	Rare potential Loss-of-Function (LoF) variants unveil genes affected by genetic knockout	61
4.1.4	Genetic burden analysis reveals new targets with increased genetic load	62
4.1.5	Edgetics provides insights to the effect of genetic variants on PPI networks	63
4.1.6	Integrative analysis and structural variation	64
4.1.7	Variant distribution and frequency of deleterious events	67
4.1.8	Enrichment Analysis	68
5	The neuronal cytoskeleton - a common feature	75

5.1	Background: Neurodevelopment and the neuronal cytoskeleton	75
5.1.1	The microtubule cytoskeleton, structural element and track for neurotransmitter traffic	77
5.1.2	Neurofilaments provide stability to mature axons	77
5.1.3	Actin microfilaments play a critical role in dendritic spine modeling	78
5.2	The neuronal cytoskeleton and schizophrenia	78
5.2.1	Microtubule stability, dynamics and the microtubule-organizing center	78
5.2.2	Microtubule-based transport	79
5.2.3	Neuronal migration	81
5.2.4	Actin-based processes	81
5.2.5	Cases with disrupted neuronal cytoskeleton	82
6	Technical Discussion	85
6.1	Data quality	85
6.2	PCR duplicates - PCR effect or reading same the fragment twice	85
6.3	Phasing strategy and its impact on variant calling	86
6.4	The influence of sample size on research conclusions	88
6.5	Association vs. Causation	89
6.6	Candidate genes in schizophrenia	90
7	Conclusion and Outlook	93
7.1	Schizophrenia and the neuronal cytoskeleton	93
7.2	Novelty of the work and contribution to the field of schizophrenia	94
7.3	Limitations and future directions	94
7.4	Conclusion	97
8	Appendix	99
9	Further projects and publications	105
9.1	KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis	106
9.2	Effect of Epstein-Barr viral miRNAs on $CD4^+$ and $CD8^+$ T cells	110
9.3	DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences	136
	List of Tables	141
	List of Figures	143
	List of Abbreviations	149
	Bibliography	151

INTRODUCTION

1.1 Motivation

In 2015, Ralph Adolphs published an article about 23 unsolved problems of neuroscience [2]. His list includes several distinct groups of problems that are ordered by their solvability. Some, like the questions how single neurons compute or how sensory transduction works, might be solved within few years. Others, like how learning and memory works or how the brain represents abstract ideas, will most likely take a much longer time, if ever solved. Within the same category, Adolphs also lists the causes of psychiatric and neurological illnesses like schizophrenia. Until this day, complex mental diseases remain an enormous challenge for affected individuals as well as researchers. This does not only include the causes of the disease, but also related aspects such as individualised treatment or even the definition of the disease itself. Adolphs lists the question about how we could cure psychiatric and neurological diseases as a goal that might not even be solved within the next 50 years.

The situation requires a joint effort by researchers and physicians as, according to the World Health Organization, schizophrenia affects more than 21 million people worldwide with approximately only half of them receiving care for their condition.

The main reason for this situation, and the most prominent challenge that remains in the field of schizophrenia, is the incomplete understanding of the underlying mechanisms and causes of this disorder. Especially for an intended individualized and effective treatment, it is inevitable to understand the exact biological processes that lead to this specific phenotype.

Epidemiological studies showed a high pathogenetic relevance of genetic factors [3] in schizophrenia and more recent twin studies estimated the heritability of schizophrenia to be almost at 80% [4]. Although these findings as well as a certain degree of environmental influences are well established, most of the genetic factors remain to be discovered.

While in the recent years much insight could be gathered by Genome-wide association studies (GWAS) studies, they are still limited to associations with no direct link towards causation. Due to the nature of GWAS studies, the association between a genetic marker and a disease can result from a causal variant in Linkage Disequilibrium (LD). In this case, the marker is a biological indicator that is not necessarily mechanistically linked to the studied phenotype.

The currently most promising approaches are the large scale sequencing efforts that have become feasible during recent advances in sequencing technologies. Unlike GWAS studies, sequencing allows to reveal functional disease relevant genetic variants and move on from association or statistical correlation towards causation.

In this dissertation, I will examine the presence and impact of functional genetic variants with clear disease-related context using a data set of well-selected schizophrenia trios and additional individual patients. By applying state-of-the-art software and a unique combination of computational methods that cover various aspects of the mutational spectrum like *de novo*, Loss-of-Function (LoF) and compound heterozygous variants, I aim to discover a comprehensive, high quality set of candidate genes. Throughout manual curation, I will investigate the potential genetic candidates and put them into the context of current schizophrenia research. Based on the gained insights, I will discuss the findings that will potentially contribute to the understanding of this complex disease.

1.2 Mental diseases, a global burden

The term "mental disease", also called "mental illness" or "psychiatric disorder", describes a condition that comprises a broad range of different symptoms and typically has an impact on cognition, emotion, and behaviour of affected individuals. Such disorders include diseases like depression, bipolar affective disorder, schizophrenia and other psychoses, dementia, intellectual disabilities and developmental disorders including autism [5]. Taken together, these diseases present a major burden for global health.

In 2014, there were an estimated number of 43.6 million adults aged 18 or older in the United States with any mental illness [6]. This number represents around 18.1% of the total U.S. adult population. This also includes 9.8 million adults with any serious mental illness, that is defined as any mental, behavioural, or emotional disorder that substantially interferes with or limits one or more major life activities [6]. The true global burden of mental diseases might still be underestimated due to various reasons like the overlap between psychiatric and neurological disorders [7]. Mood disorders, including major depression, dysthymic disorder and bipolar disorder, are also the third most common cause of hospitalization in the U.S. for both youth and adults aged 18-44 [6]. Individuals living with such serious mental diseases typically face an increased risk of having chronic medical conditions as well as a reduced life expectancy by 25 years [6].

A recent review and meta-analysis about the prevalence of common mental disorders showed

that about one in five adults experienced a common mental disorder within the past 12 months [8]. Although such prevalence estimates vary widely, it is assumed that the large spectrum of diverse diseases affects over a third of the population in most countries during their lifetime [8]. Despite these high rates, the pathogenic mechanisms of psychiatric disorders are largely unknown. Still, it is commonly accepted that they are caused by various contributing factors. This includes mostly neurochemistry, brain structure, and genetics but also environmental effects like lifestyle and the interaction between these.

Although mental diseases are commonly seen as distinct disorders, an analysis of the five disorders in the Psychiatric Genomics Consortium (autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia) revealed genetic risk loci with shared effects on all of those five major psychiatric disorders [9]. The authors analysed data for the five disorders in 33,332 cases and 27,888 controls of European ancestry. Especially calcium-channel activity genes were proposed to have pleiotropic effects on psychopathology, loci which were also previously associated with bipolar disorder and schizophrenia [9]. Although findings of such cross-disorder effects are especially helpful for investigating the common co-occurrences of clinical phenotypes in individual patients, they are inherently based on the assumption that mental disorders can be distinguished diagnostically from one another without or with few errors. However this, in fact, is a common and well-known problem [10]. A 10-year longitudinal study of 470 first admission patients with psychotic disorders indicated a high risk of misclassification at early disease stages [11]. About 50% of the diagnoses changed at some point during the study. The largest proportion of diagnostic shifts was observed for schizophrenia. Out of 306 individuals with a non-schizophrenia diagnosis, 98 were later on diagnosed with schizophrenia. 89.2% of participants that were initially diagnosed with schizophrenia retained the diagnosis at year 10. In most cases, the change of the clinical picture explained the shift of diagnoses. While poorer functioning and greater negative and psychotic symptoms predicted a subsequent shift to schizophrenia, better functioning and lower negative and depressive symptoms predicted the shift to bipolar disorder [11]. Overall, these findings draw a concerning picture of misclassification in mental disorders with serious implications for patient treatment as well as research [11].

Although the situation for people suffering from mental health problems undeniably has improved during the last decades, there are still attitudes within most societies that view symptoms of psychopathology as threatening and uncomfortable. These attitudes frequently foster stigma and discrimination towards affected people. Further work is necessary to uncover the remaining missing pieces in genetics, pathophysiology and pharmacology and to ultimately lift the global burden of mental diseases.

adversity, etc. Presumably this process will accelerate, and the term schizophrenia will be confined to history, like "dropsy". [19].

As a consequence the "definition" of schizophrenia that is used during my work refers to concept of a psychosis spectrum disorder instead of a clearly defined and distinct disease.

1.3.2 Classification, symptoms and subtypes

Unsurprisingly, the classification of schizophrenia is, so far, an unsolved problem. Although multiple methods and ideas exist, the complex manner and different phenotypes of the disease lead to inevitable misclassifications. Typically schizophrenia is diagnosed and classified using a set of occurring symptoms that, to a certain extent, define this mental disorder.

1.3.2.1 Positive, negative, and cognitive symptoms.

According to the National Institute of Mental Health (NIH), schizophrenia symptoms are divided into three main categories; positive, negative, and cognitive symptoms. Each category includes sets of specific types of symptoms that each affect different parts of everyday life.

Positive symptoms are generally not observed in healthy people and may have different forms depending on the individual phases of the disease and medication. Most common positive symptoms include but are not limited to the following:

- **Hallucinations** are typically defined as sensory experiences (vision, hearing, smell, taste, or touch) that occur in the absence of a specific stimulus. They occur in many psychotic disorders, bipolar disorder and major depression with auditory hallucinations being the most common type in schizophrenia [21].
- **Delusions** are false personal beliefs that are not related to a reason or contradictory evidence and are not explained by a person's usual cultural and religious concepts. Delusions may include thoughts about external mind/behaviour control or cause seeing supposedly hidden messages in television and newspapers [21].
- **Thought disorder** describes incomprehensible written and spoken language and can be illustrated as the so called "thought blocking", where patients suddenly interrupt sentences while talking before completion. Affected patients may experience this as if their thoughts have been taken out of his or her head [21].
- **Movement disorders** is characterized by extreme body behaviour in a sense that patients might repeat certain motions several times or, otherwise, become catatonic, that is defined as a state in which the individual wont respond to others and stay motionless. It should also be noted that there is still an ongoing debate about to which extent observed movement abnormalities are reflecting the disease process of schizophrenia or are rather a result of the antipsychotic treatment [22].

Negative symptoms, in contrast to positive symptoms, are deficits of normal behaviour like a lack of emotion or social withdrawal which can also be caused by other reasons than schizophrenia. As a result of this, it is not always possible to assign those conditions to schizophrenia. People suffering from such negative symptoms might appear lazy or unwilling to help themselves and typically require help with everyday tasks and routines. Negative symptoms can include reduced expression of emotions, reduced feelings of pleasure in life, difficulty beginning and sustaining activities or reduced speaking [21].

Cognitive symptoms, like negative symptoms, tend to be difficult to be recognised as part of the disorder. Such symptoms include poor "executive functioning", which is the ability to use and process information, or trouble with focusing attention. Consequences of such effects can cause great limitations in normal life for affected patients as they are also related to worse employment and social outcomes [23].

1.3.2.2 The schizophrenia spectrum

In the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM) (DSM5), the term "Schizophrenia Spectrum" was introduced [24]. This term, as well as the proposed classification scheme, show that schizophrenia is rather understood as a collection of several symptoms than a strictly defined phenotype (See also Section 1.3.1). Although the separation into distinct subtypes was dropped in this newest version due to their limited clinical utility, scientific validity and reliability [25], it is clear that the disorder 'schizophrenia' includes phenotypes of different forms [20]. An intuitive visualization is shown in Figure 1.1. It depicts the range of several mental disorders starting from mental retardation up to bipolar disorder, that differ by their manifestation of certain symptoms. Consequently it is assumed that the disorders, to a certain extend, share a common genetic background [20].

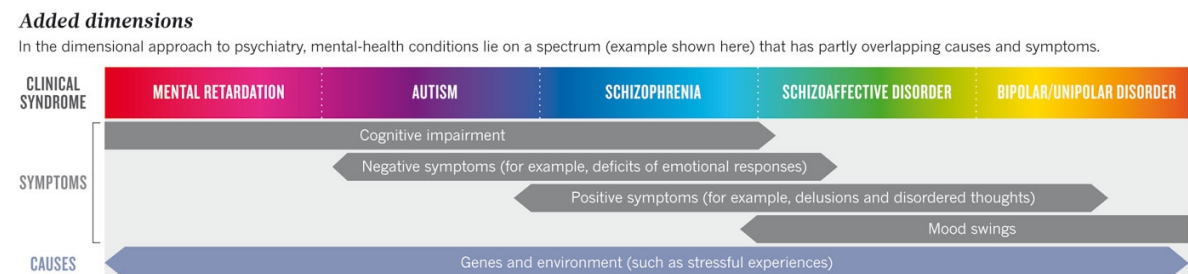


Figure 1.1: Spectrum of mental disorders. Various clinical syndromes are closely related by the degree of manifestation of certain symptoms. A common genetic background is shared by all syndromes. Figure taken from [20]

1.3.2.3 Schizophrenia Subtypes

Even though schizophrenia subtypes have been excluded from the latest DSM version [24], they have been used for extended period of time and can still be found in recent literature [26–28]. Therefore I will briefly describe the most commonly used classifications of schizophrenia, which are typically referred as schizophrenia subtypes, in the following sections.

Catatonic schizophrenia describes a clinical picture mainly characterised by difficulties in movement. Affected individuals may be unable to speak or move and resist instructions or attempts to be moved. In contrast, patients suffering from catatonic schizophrenia may also behave in extreme and hyperactive ways, known as catatonic excitement. Today, the manifestation of catatonic schizophrenia has become rare due to improved pharmacological treatment methods [29, 30].

Paranoid Schizophrenia may be the best known and most common subtype of schizophrenia. Affected individuals mainly suffer from delusions and auditory hallucinations. Delusions are often related to increased aggressiveness and violence. Hallucinations are mostly characterised by hearing voices criticising the individuals faults. As a result, patients may lose touch with reality but in comparison to other subtypes of schizophrenia, this subtype may be a less severe one, as the ability to lead an almost normal life can still be given, depending on the symptoms intensity. However paranoid schizophrenia is still a serious illness which causes lifelong complications. Like catatonic schizophrenia, the possibility for improvement through antipsychotic treatment is given [29, 30].

Disorganized schizophrenia typically describes individuals that suffer from disorganized thinking. The term disorganized behaviour can include unstructured thinking and previously mentioned "thought blocking". Also a lack of emotion is common for disorganized schizophrenia. Symptoms can be so severe for the affected individuals that they are unable to perform normal life activities. This type of schizophrenia usually occurs between the ages 15 and 25 years [29, 30].

Undifferentiated and residual schizophrenia can be seen as collective terms to include individuals which don't fall into the first three subtypes. Individuals diagnosed with undifferentiated schizophrenia typically show fluctuations of different symptoms or suffer from symptoms that do not fit into the three main subtypes. An example for this group of schizophrenia might be the Scottish DISC1 family in which a chromosomal translocation was found to segregate with mental illness including schizophrenia and schizoaffective disorder [31]. This and subsequent work showed that DISC1 in predisposing individuals to a wide range of psychiatric conditions, psychological traits, and biological phenotypes [31]. In contrast, residual schizophrenia describes individuals showing no prominent symptoms or only to a small extend. Nonetheless residual

schizophrenia patients are still affected by other schizophrenic symptoms although their manifestations are mostly diminished. [29, 30].

1.3.3 Pathophysiology

Theories that try to explain the pathophysiology of schizophrenia are mostly based on abnormalities in neurotransmission. Several hypothesis, such as the dopamine or glutamate hypothesis of schizophrenia have been proposed to explain these underlying mechanisms. However, study results are often contradictory. Even fundamental characteristics like whether schizophrenia should be considered as a predominantly neurodevelopmental or neurodegenerative disorder are still point of discussion [32]. In the following, I will present common models that try to explain the pathophysiology of schizophrenia.

1.3.3.1 A neurodevelopmental model of schizophrenia

The central assumption of the neurodevelopmental model of schizophrenia is that the mental status is the end stage of abnormal neurodevelopmental processes that began years before the onset of the illness [33]. As shown in Figure 1.2, normal cortical development involves proliferation, migration, arborization (circuit formation) and myelination, each occurring at different time stages. Cortical development is also characterized by a progressive reduction of grey-matter volume, mostly due to the pruning of the neuronal arbor and myelin deposition, and an increase of inhibitory and a decrease of excitatory prefrontal synapses during adolescence [33]. In this neurodevelopmental model of schizophrenia, it is assumed that a reduced interneuron activity and excessive excitatory pruning together with a myelination deficiency lead to an altered excitatory-inhibitory balance and reduced connectivity in the prefrontal cortex [33].

1.3.3.2 A neurodegenerative picture of schizophrenia

In contrast to the neurodevelopmental model of schizophrenia, the neurodegenerative model assumes an abnormal neurodegeneration as the cause of schizophrenia [34]. Although the results of schizophrenia studies mostly argue against neurodegeneration, there still exists support for the degeneration in certain groups of patients with schizophrenia. Besides subcellular biochemical evidence in the form of oxidative damage, also genetic defects in *bcl-2*, which regulates apoptosis, have been found in schizophrenia [34].

Also the glutamate hypothesis of schizophrenia includes glutamate as a factor in neurodegenerative processes since it is known to be a major factor programmed cell death. A number of in-vitro studies showed that high concentrations of glutamate can act as a neurotoxin which can lead to neuron apoptosis [35]. High glutamate concentrations in the brain have also been associated with known neurodegenerative disorders like Alzheimer's disease. Therefore the hypothesis has been made that apoptosis may contribute to the pathophysiology of schizophrenia [36].

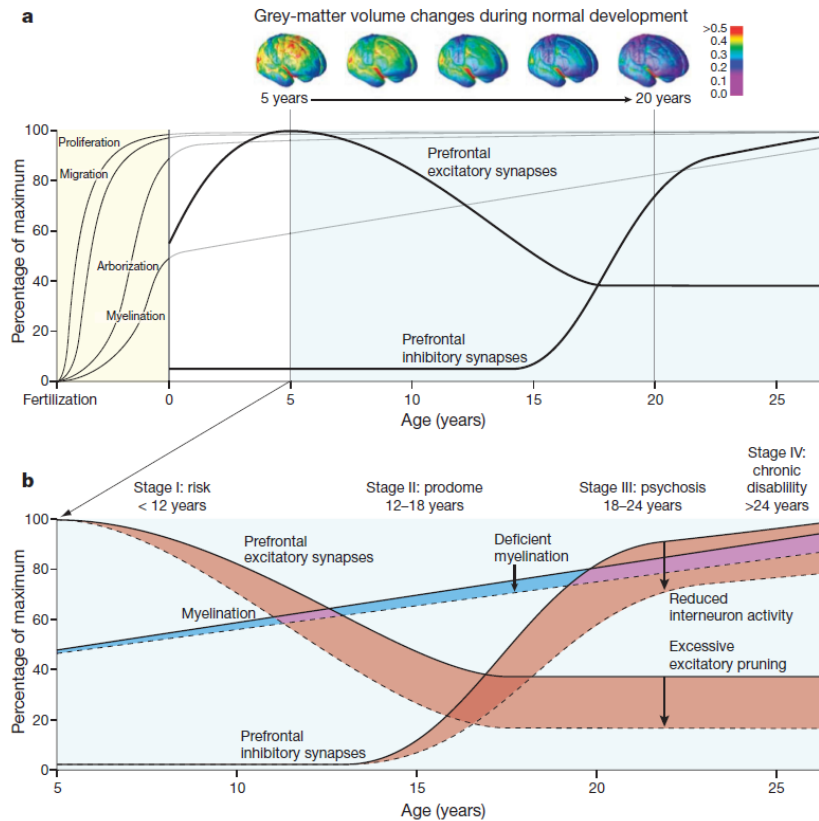


Figure 1.2: **(a)** Normal cortical development including proliferation, migration, arborization (circuit formation) and myelination. **(b)** Reduced interneuron activity and excessive excitatory pruning together with a myelination deficiency lead to an altered excitatory-inhibitory balance and reduced connectivity in the prefrontal cortex. Figure taken from [33]

1.3.3.3 The dopamine hypothesis of schizophrenia

Dopamine is the most extensively investigated neurotransmitter in schizophrenia. The dopamine hypothesis originated from the observation that drugs antagonizing dopamine were found to be effective in the treatment of schizophrenia. Interestingly, dopamine has been implicated in mediating aberrations in developmental processes such as neuronal proliferation and migration as well as pruning, and degenerative processes such as oxidative stress and excitotoxicity [34]. The dopamine hypothesis of schizophrenia is based on abnormal levels of dopamine in different parts of the human brain. Low prefrontal dopamine activity which causes negative symptoms, leads to an increase of dopamine activity in mesolimbic dopamine neurons. This condition then leads to the typical schizophrenic positive symptoms [37]. It is also known that the prefrontal dopaminergic transmission is mainly performed by D_1 receptors and that a dysfunction of these receptors causes cognitive impairment and negative symptoms [38].

The D_1 receptors are one of several subtypes of the dopamine receptors. They are implicated in many neurological processes, including motor behavior, motivation, and working memory

and therefore play major roles in neuropsychiatric disorders [39]. Studies about the levels of D₁ receptors in the prefrontal cortex using radiotracers, mainly found a decreased number of receptors in untreated schizophrenic patients while a study performed in 2002 using a different radiotracer showed increased D₁ levels [38]. One possible interpretation given in this study would support the dopamine hypothesis as the increased D₁ levels can be seen as a compensatory response to deficits in presynaptic dopamine functions [40]. Evidence for the dopamine hypothesis has also been seen in the mechanisms of many neuroleptics which bind to D₂ receptors and therefore decrease the dopamine transmission. As a result, an improvement concerning the positive symptoms can be seen [38]. Although results of several studies confirmed and supported the dopamine hypothesis of schizophrenia, still inconsistent statements from different studies can be found. This indicates that there are still uncertainties concerning the role of dopamine in schizophrenia.

1.3.4 Etiology

After many years of research, the exact causes of schizophrenia still remain unclear. Although it is widely accepted that the various phenotypes of this disease are caused by multiple factors, only a few have been proven. Early environment, neurobiology, and psychological and social processes appear to be important contributory factors [21]. Also some recreational and prescription drugs seem to cause or worsen symptoms [21].

One explanation for the development of the disease are complications during pregnancy. Especially fetal disturbances during the second trimester have been linked to a doubled risk for the offspring developing schizophrenia [42]. Also gestational diabetes, emergency cesarean section and low birth weight are assumed to contribute to an increased lifetime risk [42].

Arguably the most scientific evidence exists for genetic factors as the primary causation of schizophrenia. A recent nationwide Danish twin study [4] based on 31,524 twin pairs including 448 that were affected by schizophrenia, confirmed the heritability at 79%. This finding is in accordance with estimates from previous twin studies of 83% and 82% [18, 43] as well as a meta-analysis of 12 twin studies of 81% [44]. The investigation of phenotypic concordance rates between various degrees of relatedness (Figure 1.3) demonstrate the range between monozygotic twins and unrelated spouses. While the lower end of 1% is close to the overall population prevalence, the estimations for monozygotic (33% - 44.3%) and dizygotic (7% - 12.1 %) twins vary among studies [4, 41]. Although those numbers indicate a significant contribution of genetics, it also hints that the illness vulnerability is also partially influenced by various other factors. Still, the low concordance rate of monozygotic twins does not imply a low contribution of genetic factors in disease liability. Instead, these results can be interpreted as an indication that environmental interactions might sometimes be required to trigger the actual disease phenotype [45]. Such environmental events that already have been linked to schizophrenia, include childhood trauma, minority ethnicity, residence in urban areas and social isolation [42].

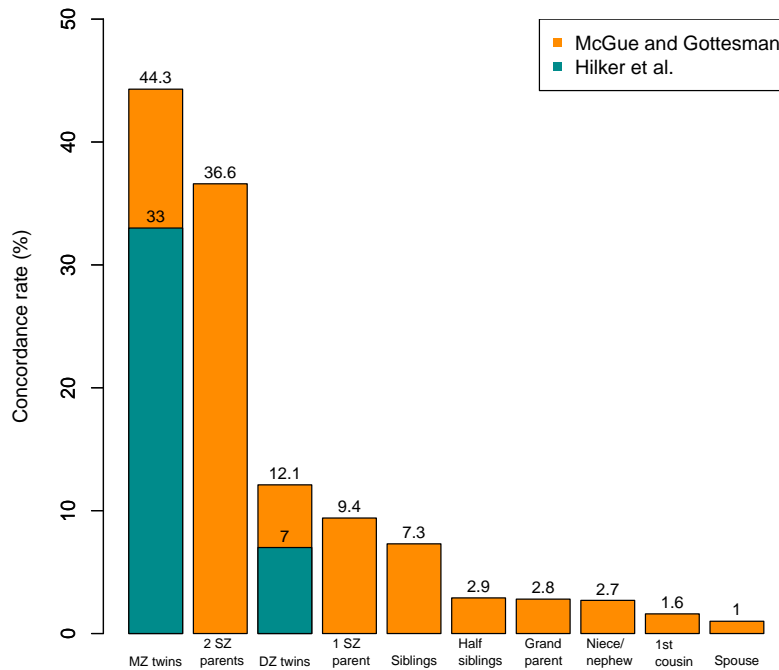


Figure 1.3: Concordance rates for different degrees of relatedness. With decreasing relatedness, the genetic concordance rates drops from 33% -44.3% to 1%. MZ, monozygotic; DZ, dizygotic; Data taken from [4] and [41]

1.3.5 Genetics of schizophrenia

Similar to the disease itself, also the genetics of schizophrenia is highly complex and heterogeneous. Most of the genetic risk to schizophrenia is suggested to be due to multiple interacting loci, each causing a small increase in risk as shown by family and twin data [46]. In contrast, genome-wide linkage scans for one or a few common major gene effects mostly fell short of genome-wide significance [47].

Today, it is commonly assumed that schizophrenia is a polygenic disease with a mutation frequency spectrum that comprises a mix of common genetic variants of only small individual effect as well as rare, highly penetrant genetic variants of larger effects [48]. In the following sections, I will provide an overview about several key findings and commonly accepted theories in the field of schizophrenia genetics (Table 1.1).

1.3.5.1 The impact of *de novo* mutations in schizophrenia

Major developments in sequencing technology have taken place in the last years enabling rapid and increasingly economical whole-exome or whole-genome sequencing. These sequencing methods can e.g. be applied for case control studies or for the detection of *de novo* mutations by sequencing both unaffected parents and the affected child. In recent years, multiple studies have provided increasing evidence for the importance of *de novo* variants in schizophrenia.

A study by Xu et al. [49] sequenced the exomes of 53 trios with sporadic cases, their unaffected parents, as well as 22 unaffected controls trios. The authors identified 40 *de novo* mutations in 27 cases affecting a total of 40 genes, including a potentially disruptive mutation in DGCR2, a gene located in the schizophrenia-predisposing 22q11.2 microdeletion region. A follow-up study by the same group sequenced 795 exomes from 231 parent-proband trios enriched for sporadic schizophrenia cases, as well as 34 unaffected trios. They observed an excess of *de novo* nonsynonymous single-nucleotide variants as well as a higher prevalence of gene-disruptive *de novo* mutations relative to controls. They also found four genes (LAMA2, DPYD, TRRAP and VPS39) affected by recurrent *de novo* events, which is unlikely to have occurred by chance [50]. Interestingly, a *de novo* mutation in another member of the laminin gene family, LAMA1 was also described in another schizophrenia sequencing study reported by Girard et al. [51]. Overall, it is commonly accepted, that *de novo* variants play a major role in the genetics of schizophrenia [52].

1.3.5.2 Shared genetic etiology with autism

As indicated in Figure 1.1, schizophrenia shares a common symptomatic basis with several related diseases like autism and schizoaffective disorder. Therefore it comes as no surprise that findings indicate partially shared genetic etiology.

McCarthy et al. [53] carried out exome sequencing on 57 trios with sporadic or familial schizophrenia. In sporadic trios, they observed a ~3.5-fold increase in the proportion of nonsense *de novo* mutations. Genes at these loci overlapped with genes implicated in autism (e.g., AUTS2, CHD8 and MECP2) and intellectual disability (e.g., HUWE1 and TRAPPC9), partially supporting a shared genetic etiology between these disorders. Functionally, CHD8, MECP2 and HUWE1 converge on epigenetic regulation of transcription suggesting that this may be an important risk mechanism. Likewise, DPYD, a candidate gene identified by Xu et al. [50], had also been found to be associated with autism [54].

Consequently, these results imply that there are specific genetic loci and alleles that increase an individual's risk of developing any of these two disorders. Additionally, these loci may represent a common biological pathway for both disorders.

Study	Data	Finding
Xu et al. (2011) [49]	53 trios with sporadic cases, 22 controls trios	40 <i>de novo</i> mutations in 27 cases affecting 40 genes; potentially disruptive mutation in the schizophrenia-predisposing 22q11.2 microdeletion region
Xu et al. (2012) [50]	231 parent-proband trios	Excess of <i>de novo</i> nonsynonymous single-nucleotide variants as well as a higher prevalence of gene-disruptive <i>de novo</i> mutations
Gulsuner et al. (2013) [55]	105 SZ cases, 84 unaffected sibs, and 210 unaffected parents	Disruptions of fetal prefrontal cortical neurogenesis may be critical to the pathophysiology of schizophrenia
Fromer et al. (2014) [56]	623 parent proband trios	<i>De novo</i> mutations overrepresented glutamatergic among postsynaptic proteins comprising activity-regulated cytoskeleton-associated protein (ARC) and N-methyl-d-aspartate receptor (NMDAR) complexes
Purcell et al. (2014) [57]	2,536 SZ cases and 2,543 controls	Polygenic burden arising from very rare mutations enriched among ARC, NMDAR, and PSD protein complexes as well as FMRP targets and calcium channel complexes
Singh et al. (2016) [58]	4,264 SZ cases, 9,343 controls, and 1,077 parent-proband trios	Genome wide significant association between schizophrenia and rare loss-of-function variant in SETD1A

Table 1.1: Several key studies investigating schizophrenia genomics.

1.3.5.3 Large scale and meta-studies

Due to the specific genetic background of schizophrenia, including mostly variants with low allele frequencies, large scale studies are required to increase chances for detecting causal sequence aberrations. In recent years, several studies gathered genetic data from several thousand schizophrenia patients and performed joint analyses at large scale.

The largest individual *de novo* trios and case-control sequencing studies of schizophrenia so far were those respectively of Fromer et al. [56], Purcell et al. [57] and Singh et al. [58]. Fromer et al. obtained further evidence for shared genetic etiology between schizophrenia and both intellectual disability and autism spectrum disorder by testing for overlap of genes affected by *de novo* loss-of-function mutations. The study shows an overlap between schizophrenia, autism spectrum disorder and intellectual disability at the resolution not just of loci or even individual genes, but even of mutations with similar functional (loss-of-function) effects. Although Fromer et al. were unable to confirm findings from some of the smaller studies that schizophrenia cases are enriched for small *de novo* mutations, they did find that this type of mutation was overrepresented among the same complexes implicated by *de novo* CNVs, specifically the ARC and NMDAR complexes. Mutations were additionally enriched in messenger RNAs which are targets of fragile X mental retardation protein (FMRP).

Purcell et al. identified a polygenic burden arising from very rare (frequency less than 1 in 1000 chromosomes) mutations and that once again, these were enriched among ARC, NMDAR, and PSD protein complexes as well as FMRP targets and calcium channel complexes. Despite their

size, neither of these studies was able to find individual genes that would achieve significance after correction for multiple testing.

However, recently, the UK10K group in collaboration with several partners, was able to do so through meta-analysis of their own case-control and previously published *de novo* mutation data sets. This study [58], which included a total of 4,264 SZ cases, 9,343 controls, and 1,077 parent-proband trios, obtained a genome wide significant association ($p=5.6 \times 10^{-9}$) between schizophrenia and a rare loss-of-function variant in SETD1A (SET Domain Containing 1A). The encoded protein, a histone methyl-transferase, has also previously been associated with various other severe developmental disorders [58]. Overall, these findings established histone methylation pathways in the pathogenesis of schizophrenia. Likewise, it showed the overlap between schizophrenia and other neurodevelopmental disorders at the level of a specific highly penetrant recurrent mutation in a single gene.

Although these and other studies point towards heterogeneous disruptions across various processes in neurotransmission and -development, results often lack confirmation and the mechanisms that underlie schizophrenia remain unresolved.

MATERIALS AND METHODS

The following chapter presents the data sets, methods and software tools that were used during analysis. Additionally, I will introduce the external resources and databases that were incorporated.

2.1 Datasets and resources

2.1.1 The m4-Schizophrenia sequencing project

The m4 project sequencing data consists out of 210 unrelated individuals with schizophrenia (74 females, 136 males) that were ascertained from the Munich area in Germany. Of this sample, 76% were of German descent and 24% were Caucasian middle Europeans. Case participants had a DSM-IV and ICD-10 diagnosis of schizophrenia with the following subtypes: paranoid 74.2%, disorganized 16.3%, catatonic 3.3% and undifferentiated 6.2%. Detailed medical and psychiatric histories were collected, including a clinical interview using the Structured Clinical Interview for DSM-IV (SCID I and SCID II) [59, 60] to evaluate lifetime Axis I and II diagnoses. Four physicians and one psychologist rated the SCID interviews, and all measurements were double-rated by a senior researcher. Exclusion criteria included a history of head injury or neurological diseases. All case participants were outpatients or stable inpatients. Further details can be found in previous reports [61].

Notably, the dataset also includes 37 trios (81% of German descent) with the index suffering from schizophrenia (14 females, 23 males). No exclusion criteria were applied, in 26 trios the parents were healthy, 10 trios contained one parent with psychiatric disease, in one trio both parents were affected by a psychiatric disease. Besides this information, no additional phenotype data was available during this study.

DNA was obtained from peripheral blood. DNA concentration was adjusted using the PicoGreen quantitation reagent (Invitrogen, Karlsruhe, Germany). Sequencing was performed by Eurofins Medigenomix GmbH and the Technical University of Munich (School of Life Sciences Weihenstephan) using Illumina HiSeq system (2x100 and 2x125 paired-end mode)

2.1.2 External reference resources

Next to the described m4-Schizophrenia sequencing data, I used further external resources for obtaining variant frequencies. This allows to discriminate between rare and common variants.

- **The Exome Aggregation Consortium (ExAC) [62]**

The Exome Aggregation Consortium is a coalition of several investigators that collected data from 60,706 unrelated individuals that were originally included in various disease-specific and population genetic studies. Raw data from these projects have been processed using the same pipeline and jointly variant-called to increase consistency across projects. So far, global and sub-population allele frequencies are made available.

- **The NHLBI GO Exome Sequencing Project (ESP) [63]**

The Exome Sequencing Project currently contains around 6,500 exomes (ESP6500). They were initially collected with the main goal to discover novel genes and mechanisms contributing to heart, lung and blood disorders. The ESP contains the protein coding regions of the human genome of several well-phenotyped populations.

- **The 1000 Genomes Project (1000G) [64]**

The 1000 Genomes Project provides a comprehensive description of common human genetic variation through whole-genome and deep whole-exome sequencing. Throughout multiple phases, the 1000G project has collected 2,504 individuals from 26 populations. In total, this includes more than 88 million variants (84.7 million SNPs, 3.6 million Insertions and deletions (*Indels*), and 60,000 structural variants).

2.2 Variant calling and phasing pipeline

Comprehensive analysis of Next Generation Sequencing (*NGS*) data requires the application of several different tools and methods. In the following, I will present all steps that were performed to obtain a high quality set of phased genetic variants (Figure 2.1).

2.2.1 Data preprocessing

Initial quality control of raw sequencing data was performed using a modified version of FastQC v0.10.1 [68], a quality control tool for high throughput sequence data. This step is necessary as raw sequencing data may contain severe technical artifacts or low quality reads that can

2.2. VARIANT CALLING AND PHASING PIPELINE

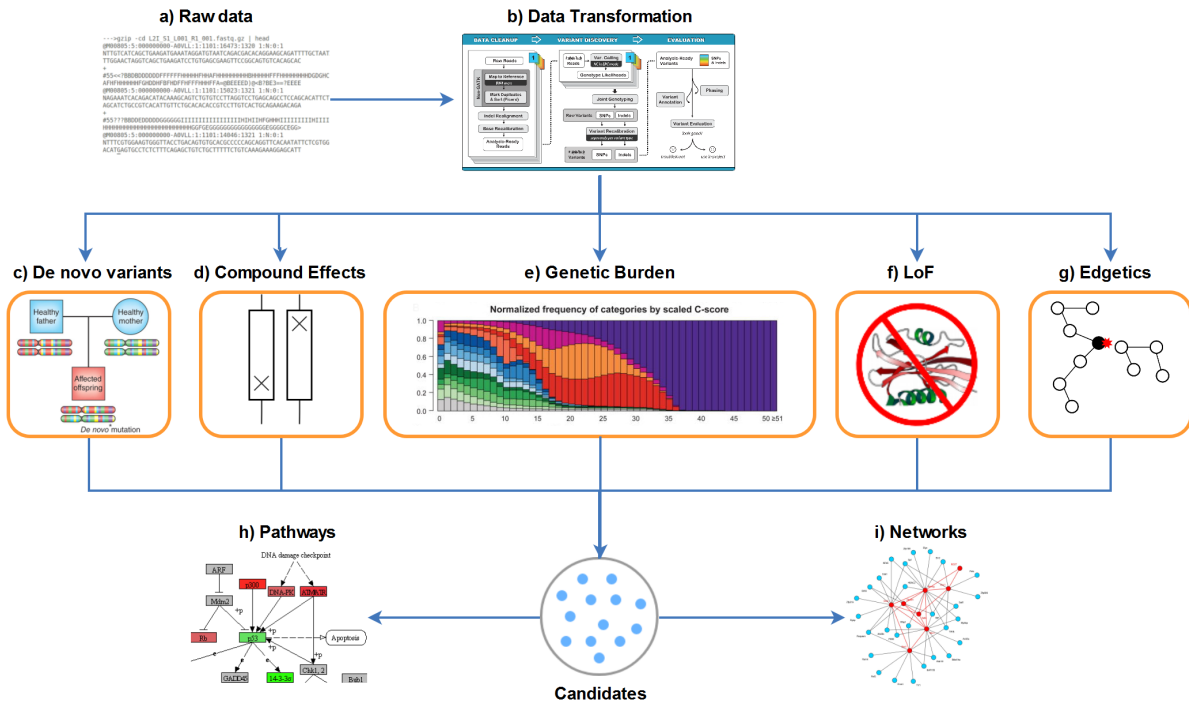


Figure 2.1: Representation of the variant processing and filtering workflow according to the applied methods. After initial quality filtering of raw data (a), reads are processed based on GATK Best Practices (b). The analysis-ready variants are then investigated throughout five distinct approaches (c-g). This results in a set of candidate genes that are further analysed by system biological methods (h-i). Subfigures b, c, e were taken from [65–67].

affect analysis outcome. Besides, Illumina reads also tend to show decreased base quality scores towards the end of sequence reads, as shown in Figure 2.2. Here, a clear tendency can be seen as the base quality drops with the increasing position index. Low base quality scores can indicate incorrect base calls by the sequencer which subsequently would lead to errors in following steps and therefore should be removed.

In order to cope with quality issues, I applied the RawReadManipulator [68], an in-house read filtration tool, for trimming and discarding reads based on raw sequencing base quality. The RawReadManipulator makes use of the data-independence that reads (or read-pairs) have towards each other and includes most common filter processes. Here, reads were required to have a minimal mean quality of 10 and contain no 'N' bases as this might indicate technical issues during sequencing. Additionally, the read 3' ends were trimmed until the reads last base meets the required quality score threshold of 15. FastQC was then applied on the filtered reads to confirm and validate the filtering effect.

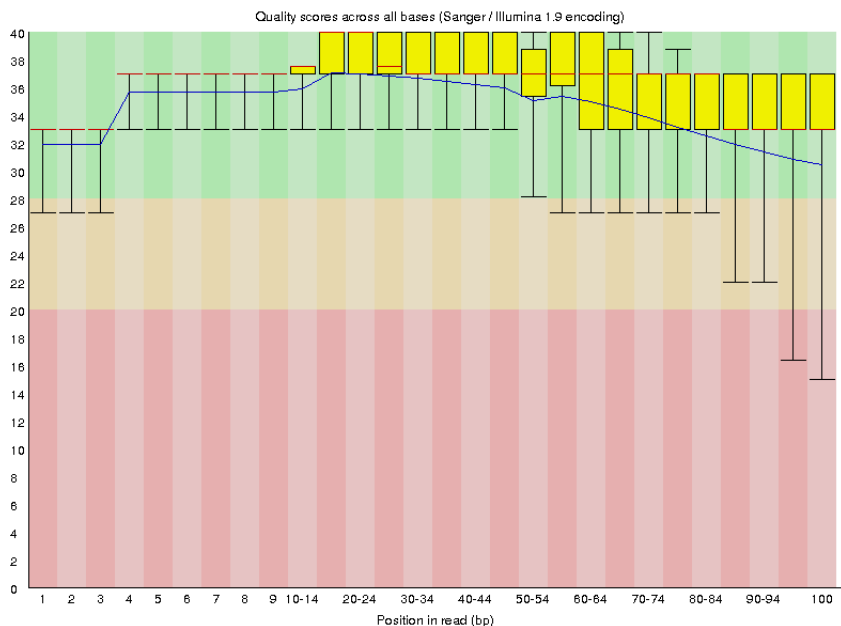


Figure 2.2: Exemplary raw quality profile showing the base quality per position index. Each bin indicates the base quality for one or multiple positions within a sequence read calculated across all available sequencing reads. Yellow boxes indicate the 25%-75% quantile with the red line indicating the median. The black whiskers show the range between the 10% and 90% quantile, the continuous blue line represents the overall mean quality.

2.2.2 Read mapping and alignment processing

After filtering, the remaining reads were mapped against the UCSC human reference genome (build hg19). The mapping was done using the Burrows-Wheeler Aligner (BWA 0.7.12-r1039) [69], a read alignment package which is based on the Burrows-Wheeler Transform (BWT) (See Figure 2.3 for further details).

Afterwards, the mapped reads are converted into Binary Alignment/Map (BAM) format using Picard [73]. I followed the Broad Institute 'Best Practices' workflow [72, 74] in order to further improve data quality and performed local realignment and quality score recalibration using the Genome Analysis Toolkit (GATK, version 3.3-0) [75].

Local realignment (Figure 2.4) aims to reduce errors introduced during the previous mapping step as especially *Indels*, read ends and homopolymer runs tend to be problematic during alignment. If mismapping occurs, this can subsequently lead to false-positive variant calls. In order to reduce this effect, a local realignment around known *Indels* is performed. During this realignment step, all reads are considered and used to create the best alternative arrangement. If this new alignment is sufficiently better than the original alignment, the proposed realignment is accepted. Here, I used the 1000G Phase 1 *Indels* and the Mills and 1000G gold standard *Indels* as reference and starting point of the realignment.

Transformation				
Input	All rotations	Sorted into lexical order	Taking last column	Output last column
<code>^BANANA </code>	<code>^BANANA </code> <code> ^BANANA</code> <code>A ^BANAN</code> <code>NA ^BANA</code> <code>ANA ^BAN</code> <code>NANA ^BA</code> <code>ANANA ^B</code> <code>BANANA ^</code>	<code>ANANA ^B</code> <code>ANA ^BAN</code> <code>A ^BANAN</code> <code>BANANA ^</code> <code>NANA ^BA</code> <code>NA ^BANA</code> <code>^BANANA </code> <code> ^BANANA</code>	<code>ANANA ^B</code> <code>ANA ^BAN</code> <code>A ^BANAN</code> <code>BANANA ^</code> <code>NANA ^BA</code> <code>NA ^BANA</code> <code>^BANANA </code> <code> ^BANANA</code>	<code>BNN^AA A</code>

Figure 2.3: Example of a Burrows-Wheeler transformation that allows efficient and reversible compression of character strings. Given a specific character combination, the BWT first generates all possible word rotations and then sorts them in lexical order. Afterwards, only the last column of the aligned character strings is stored. Due to the runs of repeated characters it can be stored efficiently and is still sufficient to reproduce the original input. The Burrows-Wheeler Aligner is based on this principle and uses it for fast and accurate short read alignment. Figure taken from [70]

The final alignment processing step involves the recalibration of base quality scores. As shown by DePristo et al., all major sequencing technologies including Illumina, 454 and SOLiD provide inaccurate quality score estimates [72]. This is of high importance as most subsequent analyses are based on evidence estimations, calculated by using quality scores. The GATK Base Quality Score Recalibration tries to identify systematic errors as they typically correlate with basecall features (e.g. quality score, position within the read, sequence context). The error is calculated empirically using known reference variants. In the course of this study, I used the recommended 1000G Phase 1 [Indels](#), the Mills and 1000G gold standard [Indels](#) and variants from dbSNP138 as such references. The calculated error-model is then applied on the original base quality scores and used to generate the new recalibrated scores.

2.2.3 Variant calling and filtering

After raw data filtering, mapping and alignment processing, the data can now be used for calling all variants compared to the human reference genome. Here, [SNP](#) and Indel calls were made by applying the HaplotypeCaller in GVCF mode. All individual gVCFs were combined and jointly genotyped with a restriction to the targeted coding regions (SeqCap EZ Exome v3). Compared to an independent and single sample approach, the joint analysis of the complete dataset allows to calculate genotype likelihoods across all samples which increases quality and reliability.



Figure 2.4: Read mapping before and after local realignment. rs-identifiers indicate potential variant sites. Local realignment was performed for two independent data sets (1,000 Genomes Pilot 2 data mapped using MAQ [71] and HiSeq data mapped using BWA). The left panel of the plot indicates original read mapping with three possible variants (rs28782535, rs28783181 and rs28788974). After realignment around known *Indels*, only one of the original *SNPs* (rs28788974) remain. In addition, a previously hidden insertion is visible (rs34877486). In both cases, the realignment clearly improves overall mapping quality and removes false-positive variants. Figure based on [72]

In general, the HaplotypeCaller performs four main steps (Figure 2.5). It first defines 'ActiveRegions' which are characterised by the presence of significant evidence for genomic variation. Within these regions, haplotypes are determined by re-assembly of the complete region using a De Bruijn-like graph. The possible haplotypes are then aligned against the reference haplotype using the Smith-Waterman algorithm which then allows to identify candidate variant sites. During the third step, the actual haplotypes quality is determined based on the support by read data. This is done by using the PairHMM algorithm which performs a pairwise alignment of each read against each haplotype. The resulting haplotype likelihood matrix is then used for obtaining likelihoods of alleles per read for each potentially variant site. In a final step, the actual sample genotypes are assigned. For each potentially variant site Bayes' rule is applied to calculate the likelihoods of each possible genotype, and selecting the most likely. The most likely genotype is then assigned to the sample.

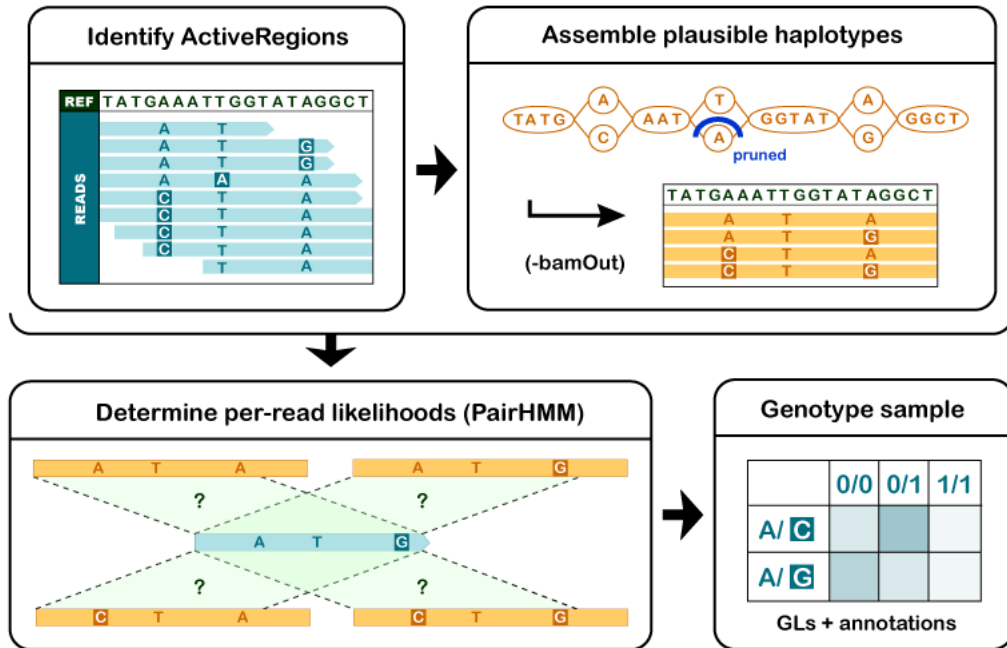


Figure 2.5: Core operations performed by HaplotypeCaller during variant calling. In a first step, regions with evidence for variants are selected (ActiveRegions). Within the ActiveRegions, all plausible haplotypes are calculated. For each haplotype, quality values are assigned by using the available read data. Finally, the most likely genotype is assigned by using the created haplotype matrix. Figure taken from [76]

Although the called variants tend to be of good quality, mutation calling algorithms tend to be very permissive which, as a result, leads to a certain number of false-positive variants in raw callsets. To remove those artifacts, I applied the [GATK Variant Quality Score Recalibration \(VQSR\)](#) that uses high-confidence known sites to determine the probability of a given variant to be true mutation. Compared to hard-filtering, this approach has the advantage to be more flexible as it is based on the sensitivity to the reference (truth) set of variants. As shown in [Figure 2.6](#), [VQSR](#) uses variant annotations to build a Gaussian mixture model and scores these variants to their relative position to the cluster. Afterwards a sensitivity cutoff can be specified and then applied when evaluating new variants out of the callset.

All discussed and analysed datasets were processed using the [GATK](#) resource bundle (HapMap v3.3, Omni v2.5, 1000G Phase 1 HC [SNPs](#), dbSNP v138, Mills and 1000G Gold Standard [Indels](#)) separately for [SNPs](#) and [Indels](#). In both cases, I used the recommended annotations for exome sequencing projects ([SNPs](#): QD, MQ, MQRankSum, ReadPosRankSum, FS and SOR; [Indels](#): QD, FS, SOR, ReadPosRankSum, MQRankSum). The tranche sensitivity thresholds were set to 99.5 ([SNPs](#)) and to 99.0 for [Indels](#).

To further improve data quality I applied additional filters as proposed by Carson et al. [77]. Genotypes were required to be supported by at least eight reads and a minimal genotype quality

of 20. At the variant level, variants were required to have an average genotype quality of at least 35 and a call rate of 88%.

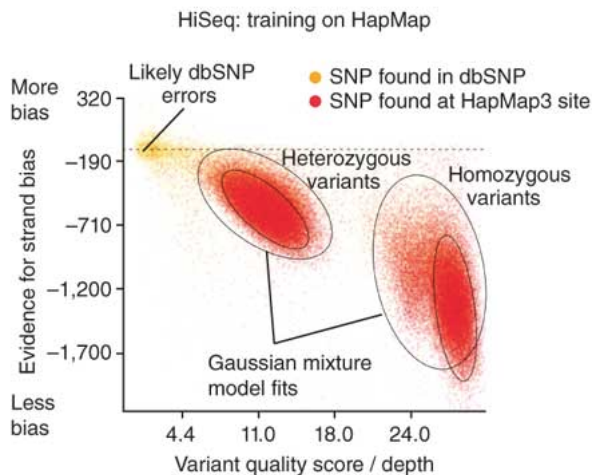


Figure 2.6: Training of the VQSR model. A gaussian mixture model is built on training data taken from dbSNP and HapMap3. In this example, the resulting model is based on two quality scores, strand bias and variant quality score. Figure taken from [72]

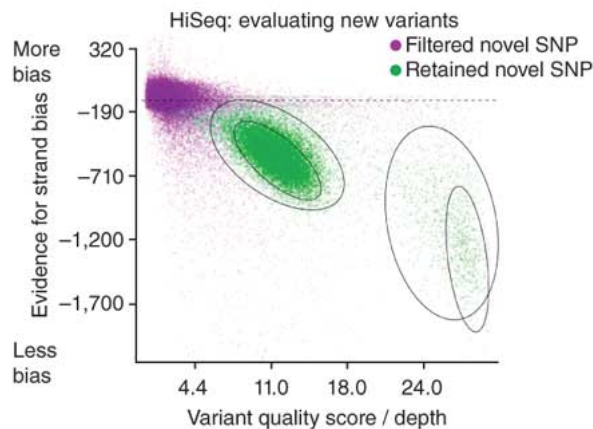


Figure 2.7: Application of the trained VQSR model. The trained model is applied on new variant data. Based on the selected stringency, variants are filtered or kept according to their relative position compared to the known variants that were used during training. Figure taken from [72]

2.2.4 Haplotype phasing

The gametic phase describes the underlying inheritance pattern and identifies alleles that are co-located on the same chromosome. This information indicates whether two specific variants have been inherited together or independent of each other. When investigating compound heterozygous variants or compound effects in general, this information is essential. Due to the different dataset types (trio vs. individual patient), I used two different approaches to generate phased genotypes. For trio data I used GATK PhaseByTransmission. Using genotype information of all family members, the tool computes the most likely genotype combination and phasing individually for each site. Sites where all samples are heterozygous remain unphased as it is not possible to determine the true genotype phase using only this type of information. As a result, the phased genotypes of heterozygous variants of the patient indicate which genotype was inherited from which parent. This information then can be used for calling compound heterozygous or *de novo* variants.

For samples without additional parental sequencing data, I used Shapeit, a fast and accurate method for estimation of haplotypes from genotype or sequencing data [78]. In contrast to the previously described approach by PhaseByTransmission which investigates each site independently, this method considers variant tuples. Based on information gained from actual read data

and/or reference panels, Shapeit decides whether both variants are located on the same allele or not. If sequence reads are available, Shapeit searches for phase-informative reads which are defined as reads that carry both variants, linking them into one haplotype. Additionally, reference panels (1000G, HapMap) that include phasing information about common variants can be used to improve phasing accuracy. In this case, information about common inheritance is extracted from large datasets that include hundreds of phased individuals. Here, I relied on both sources to gain a comprehensive set of phased genotypes. Phase-informative reads were extracted using the created [BAM](#) files and combined with the 1000 Genomes Phase 3 panel.

2.3 Variant annotation and analysis workflow

After filtering and phasing all variants, the resulting callset is prepared for a comprehensive analysis of different genetic variant types. In the following, I will present all methods that were applied to extract a reasonable set of schizophrenia candidate genes and how they were used in a joint analysis (Figure 2.1).

2.3.1 The Combined Annotation Dependent Depletion ([CADD](#)) score

The human exome typically contains thousands of genetic variants, some rare mutations cause severe genetic effects, some are considered silent. Therefore it is inevitable to rank variants according to their potential deleteriousness. Here, I used the [CADD](#) score to achieve the best possible selection of true deleterious variants. In the following, I will shortly present the underlying mechanisms and scoring scheme.

When annotating genetic variants, current methods usually focus on single variant features such as the conservation or are limited to certain variant types (e.g. missense changes). As a result, multiple annotation tools have to be applied to generate a comprehensive annotation. However, this will usually end up in a large number of independent annotation scores which all have to be considered separately. To circumvent this problem, Kircher et al. [67] invented the [CADD](#) score. The [CADD](#) score (C-Score) objectively integrates 68 annotations into a linear support vector machine and uses the trained model to score all possible substitutions. The used annotation scores include conservation metrics such as GERP [79]; regulatory information [80] like genomic regions of DNase hypersensitivity [81] and transcription factor binding [82]; transcript information like distance to exon-intron boundaries [80]; and protein-level scores like SIFT [83], and PolyPhen [84]. The obtained PHRED-scaled annotation scores express the rank of a given variant compared to all ~ 8.6 billion [SNPs](#) of the human GRCh37/hg19 reference. The higher the score, the higher the calculated deleteriousness of a given variant whereas variants at the 10th-% of [CADD](#) scores are assigned to a [CADD](#) score of 10, top 1% to a [CADD](#) score of 20, etc. The same principle can also be applied for so far unseen variants. The relationship of scaled [CADD](#) scores and categorical variant consequences are shown in the Figure 2.8.

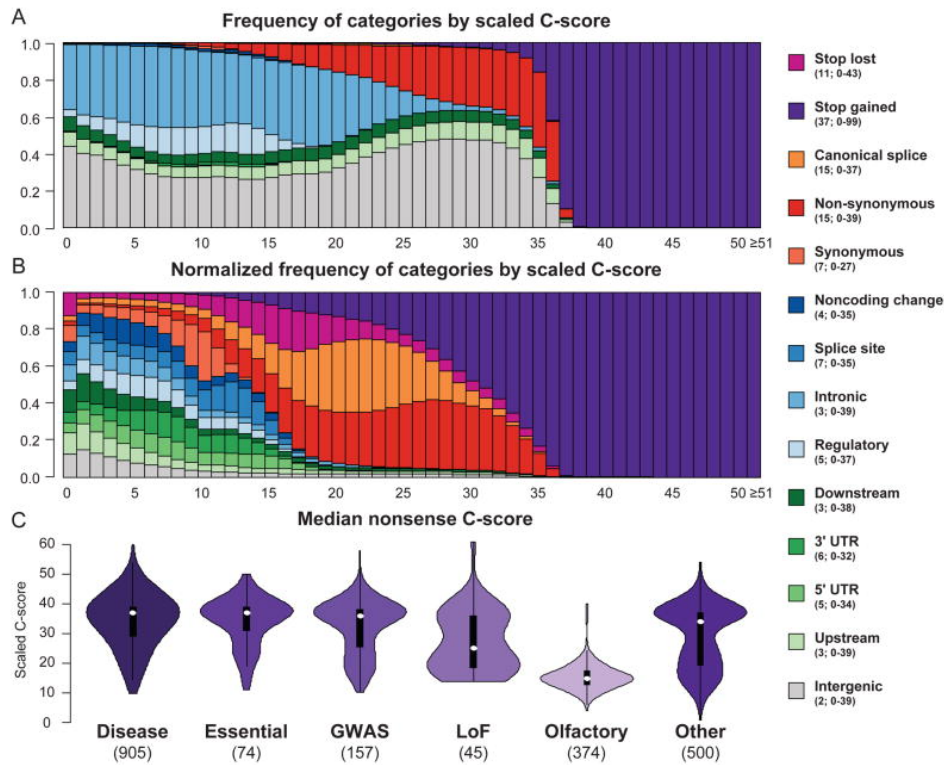


Figure 2.8: Relationship of scaled **CADD** scores and categorical variant consequences. **A** and **B** indicate the proportions of substitutions with a specific consequence for each scaled **CADD** score. **C** shows the score distribution of loss-of-function variants in specific gene sets: **Disease** genes harbor at least 5 known pathogenic mutations, **Essential** genes were predicted to be essential, **GWAS** genes harbor at least 2 loss-of-function mutations in 1000 Genomes, **Olfactory** genes encode the common olfactory receptor proteins and **Other** contains a random selection of 500 genes. Figure taken from [67]

2.3.2 *De novo* variant calling

As shown by large sequencing studies [56], *de novo* variants (Figure 2.9) strongly contribute to schizophrenia. Therefore I investigated all patients with available parental sequencing data for the presence of *de novo* mutations. As *de novo* variants typically are more error prone than normal mutations (genotypes of all three family members have to be correct), I applied the **GATK** genotype refinement workflow for improving the accuracy of genotype calls. I used the 1000G Phase 3 v4 sites for deriving the posteriors of genotype calls (b37 coordinates were converted to hg19 using Picards liftover). Putative *de novo* variants were annotated using the **GATK** VariantAnnotator. I selected all possible *de novo* mutations that were classified as high confidence by the VariantAnnotator and performed further filtering and annotation using KGGSeq (v1.0+) [85]. During the KGGSeq processing, I used several further criteria to discard potential false positive mutations. All genotypes with a sequencing depth below 8, genotyping quality below 20 or if their second smallest Phred-scaled likelihoods (PL) was below 20 were

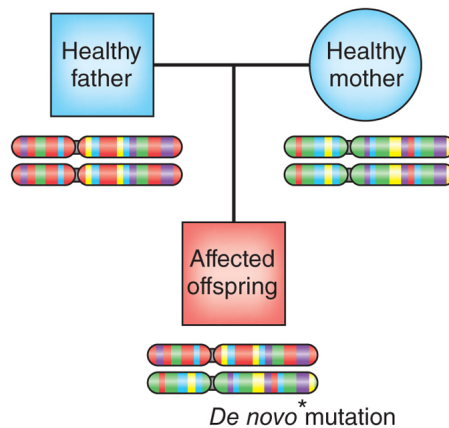


Figure 2.9: Heterozygous *de novo* mutation affects offspring and is not present in either healthy parents. Figure taken from [66].

removed. Furthermore, a genotype was ignored if the fraction of the reads carrying the alternative allele was higher or equal to 0.05 at a reference-allele homozygous genotype, lower or equal to 0.25 at a heterozygous genotype or that was lower or equal to 0.75 at an alternative-allele homozygous genotype. Additionally, variants within super-duplicated regions registered in UCSC genomicSuperDups table were also removed.

To extract potential causal *de novo* mutations, I applied a variant frequency filter of 0.05 using the 1k201304, dbsnp138, ESP6500AA, ESP6500EA and ExAC databases. All variants with annotated frequencies above the threshold were assumed to be common and therefore most likely do not contribute to the disease. In addition, I required all candidate variants to be predicted as protein sequence altering and to have a CADD (CADD v1.3) score of at least 15 to ensure a certain degree of deleteriousness.

2.3.2.1 Compound Heterozygous Variants

Although *de novo* variants strongly contribute to the genetics of schizophrenia, they are not sufficient for explaining underlying mechanisms in all cases. Therefore, I also focused on genetic compound effects like compound heterozygous variants, a combination of mutations that affect both alleles of a gene (Figure 2.10). This offers the possibility to go beyond predicting effects of single variants and instead takes into account that complex diseases are often caused by multiple genes and/or multiple mutations on individual genes [86–90].

To detect genes affected by compound heterozygous variants, I phased all samples as described in Section 2.2.4. Afterwards, I extracted all rare (ExAC MAF <0.05) deleterious (CADD \geq 15) variants. Using this set of variants, I selected all genes that carried at least a single compound heterozygous variant. Synonymous and variants for which at least one of the respective parents was homozygous for the alternate allele were discarded as they most likely do not contribute to

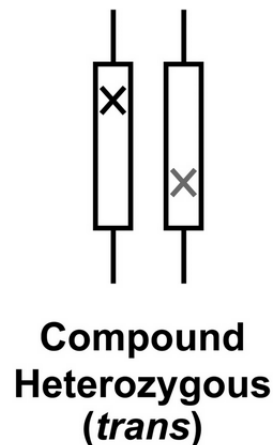


Figure 2.10: Compound heterozygous mutations (*trans*) affecting both alleles. Figure taken from [91].

the disease.

2.3.3 *De novo* gene knockouts

Dominant, homozygous and compound heterozygous **LoF** variants typically come with severe effects. Although some mutations can be classified as benign or even a gain of function, many **LoF** variants completely disrupt the function of important protein-coding genes. Here, I annotated **LoF** variants using VEP (v83) [92]. I filtered all potential **LoF** variants using the VEP HC-tag to obtain a high-quality set of **LoF** mutations. To account for frequently occurring variants, I used an ExAC frequency cutoff of 0.05 to filter common variants. I then selected all genes that were affected by a complete knock-out. This may either be due to homozygous **LoF** variants or a combination of multiple **LoF** variants. All genes that were knocked out in cases but remained functional in both parents were kept as candidates for further analysis.

2.3.4 Genes enriched for deleterious variants

Although single rare variants have can have strong effects (*de novo*, **LoF**, compound heterozygous variants), also common variants with a smaller functional impact (Figure 2.12) probably contribute to the genetics of schizophrenia. To account for this, I investigated the enrichment of deleterious variants in genes, regardless of the variant frequencies.

Genetic burden

First, I defined the genetic burden of a gene using the **CADD** score. I calculated the **CADD** deleteriousness score for all variants and used those to get the **CADD** sum scores (CSS) for each gene (Equation 2.1). I did not limit this analysis to a specific type of variants, as benign variants

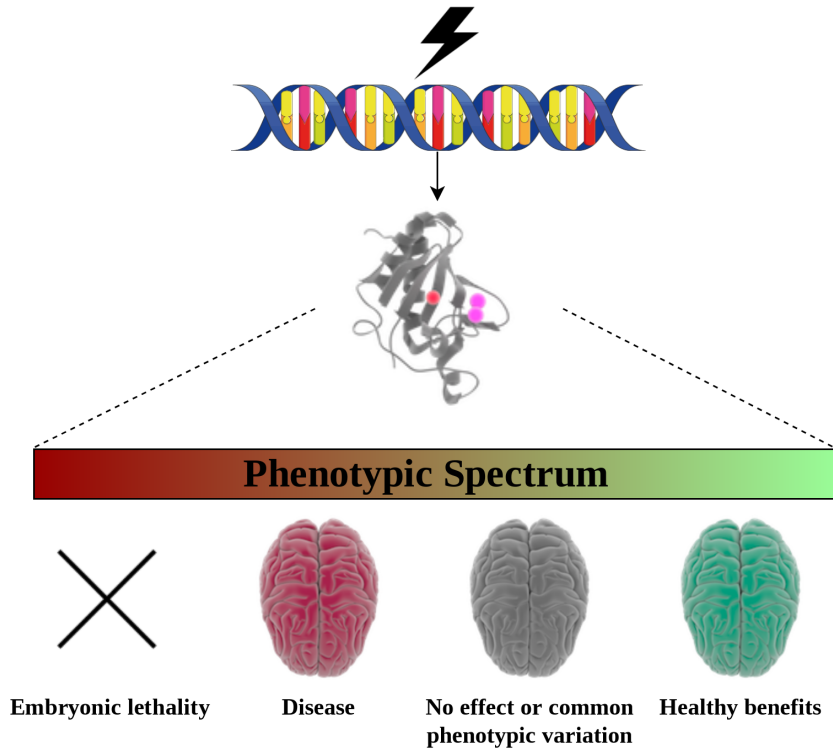


Figure 2.11: Depending on the affected gene, Loss-of-function variants can have different phenotypic effects. While the most extreme form may lead to embryonic lethality, **LoF** variants may also have healthy benefits.

are automatically down-weighted through their **CADD** scores (Figure 2.8). I then defined the CSS_{Gene} as the overall genetic burden of a gene. This values then can be compared to obtain candidate genes with unusual CSS scores.

$$(2.1) \quad CSS_{Gene} = \sum_{i=1}^n CADD - Score_i$$

Where:

n : is the number of variants that are located within the respective gene

i : is the index i-th variant within a gene

$CADD - Score_i$: is the predicted CADD score for the i-th variant

Trio comparison and gene selection

As a next step, I compared the gene-specific CSS_{Gene} within each trio in order to find patient genes with an increased genetic burden. I defined genes with an increased genetic burden as genes that matched the following criteria: $CSS_{Patient} > \max\{CSS_{Father}, CSS_{Mother}\}$. In these cases, the affected offspring harbours more, or more deleterious, variants than either of their respective parents. As this increase of the CSS (which could be due to enrichment of several mediocre

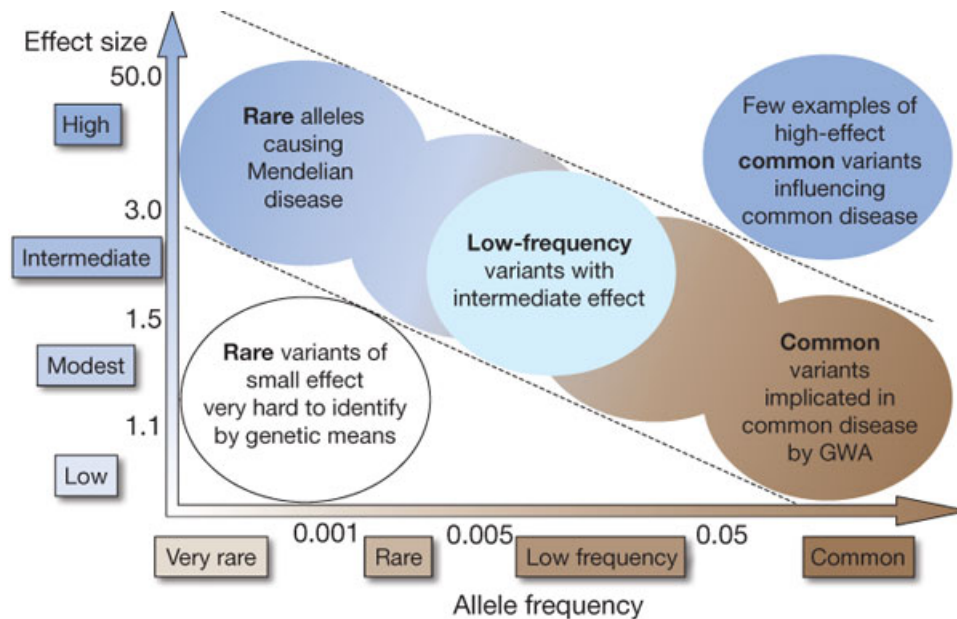


Figure 2.12: The spectrum of genetic variants ranging from common variants with low effects to rare variants with high effect sizes. Figure taken from [93]

variants or a few high impact ones) might happen by chance in a few samples, I applied an additional frequency filter to account for that. In order to be taken as a possible candidate gene, I required the gene to have an increased genetic burden in at least 1/3 of all trio-patients. This strict frequency filtering has the disadvantage that only the most extreme cases will be observed. Still, it is necessary as it provides the only reasonable way to obtain a clear set of candidates that still have the potential to carry detectable biological signals. Reducing the cutoff leads to a drastic increase of potential candidates (See Figure 3.9) which would dilute signals and hamper biological interpretation.

2.3.5 Edgetics - a link between genotype and phenotype

The term edgetics describes the analysis of protein-protein interactions and networks. By combining genotype and protein interaction information, it is possible to specify certain Edgotypes that are a result of Protein-Protein Interaction (PPI)-affecting mutations. As shown in Figure 2.13, a mutation that directly affects a PPI interaction site might have several distinct effects. Potentially it could prohibit some or all possible interactions but could also lead to novel interaction partners, depending on the actual biochemical adjustments. Compared to typical sequence-based analysis, this approach allows to go beyond common annotations and include network interaction partner information, gaining access to a broader picture of mutation effects.

Here, I used the dSysMap database (Mapping of Human disease-related mutations at the systemic level) [94] to obtain a reference list of known human protein-protein interaction interfaces.

The database contains specific amino acid sites that interact during PPIs and therefore allows to map mutations on protein structures and the human interactome. I used ANNOVAR annotation software [95] to annotate and compare the positions and amino acids with the dSysMap interfaces to extract all mutations that are located within a given PPI site. Synonymous mutations were excluded as they have no effect on the actual amino acid sequence and, as before, I relied on ExAC to discard common variants. I then selected all genes/proteins with disrupted PPI interfaces as well as the direct interaction partners as potential candidates.

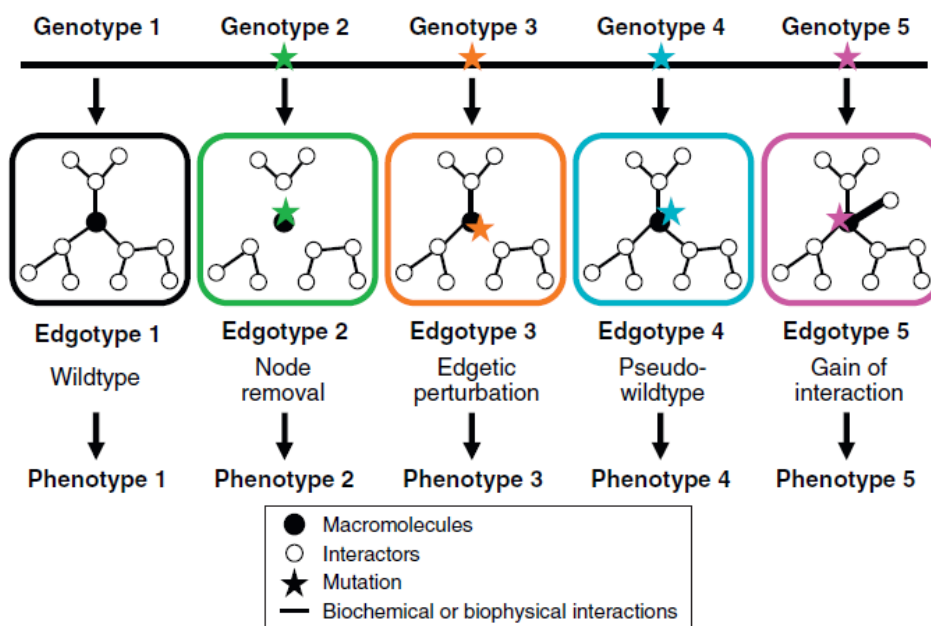


Figure 2.13: Edgotypes defined by the edgetics approach. Each mutation affecting protein-protein interactions changes the interaction network in a specific manner. Depending on the outcome, different phenotypes may arise. Taken from [96]

2.3.6 Gene set over-representation enrichment analysis

Making sense out of gene lists is not always a trivial task. One approach is the widely-used gene set enrichment analysis that checks for over-representations of certain gene groups [97]. Given a set of reference genes and annotations, the method checks whether a set of candidates is significantly enriched for a certain type of biological annotation or classification. Annotation databases that are typically used for performing this task are Gene Ontology (GO) [98] as well as biological pathways databases like KEGG [99].

In order to see whether candidate genes resulting from the five main analysis (Figure 2.1) steps are enriched for specific pathways or functionalities, I used the WEB-based GENE SeT AnaLysis Toolkit (WebGestalt) [100, 101] and the DNENRICH framework for calculating recurrence and gene-set enrichment [56] for conducting enrichment analysis.

The initial enrichment was performed using WebGestalt and a redundancy reduced functional GO database. All candidate genes that resulted from previous analyses (2.3.2 - 2.3.5) were used as input. Due to the redundancy-reduced GO database, multiple occurrences of closely related terms are avoided. Functional GO groups were required to contain at least 10 candidate genes in order to be considered. P-value correction was performed using the Benjamini-Hochberg procedure [102].

Although WebGestalt is sufficient for detecting general enrichments, it does not consider the underlying gene lengths. This can be a critical drawback as long genes can harbour more variants by chance than shorter genes, given that the overall conservation levels are similar. Therefore I additionally applied DNENRICH to confirm the previous results. In contrast to WebGestalt, DNENRICH is based on the random placement of mutations across the human exome. Throughout multiple iterations, it calculates the number of expected hits within each functional gene group.

2.3.7 Protein-protein interaction networks

PPI networks are a fundamental part for making sense out of gene lists. Using external PPI data, a given set of genes can be placed in the context of related genes or proteins, respectively.

Here I relied on the STRING database of known and predicted PPIs [103]. Interactions in STRING are mainly derived from five sources: genomic context predictions, high-throughput lab experiments, co-expression, automated textmining and previous knowledge in databases. Since not all of those sources share the same level of confidence, I restricted the known interaction sources to experiments and databases with at least high confidence (as defined by STRING). To ensure the creation of a comprehensive network, I included a maximum of 99 first shell interactors, whereas first shell interactors are defined as proteins that are directly associated with the input proteins.

2.4 Structural variants and known candidates

Besides the applied methods described in the previous sections that aim for an unbiased detection of rare SNPs and Indels, I also investigated the presence of structural variants and mutations and genes with known schizophrenia-relevance. The respective tools and approaches are described in the following.

2.4.1 Detection of long insertions and deletions using Pindel

Standard variant calling, as described in Section 2.2.3, offers the possibility to identify SNPs and short Indels, typically up to a length of 50 base pairs. Thus, a significant part of Indels can get lost. However since especially long Indels carry an increased potential to result in a complete genetic knock-out, those variants are of special interest. Due to the variants special

characteristics (variant length can exceed read length), designed variant callers are required for properly detecting long **Indels**. The tool used during this study is Pindel, a software for detecting breakpoints of large deletions, medium sized insertions, inversions, tandem duplications and other structural variants at single-based resolution from next-generation sequence data [104]. Pindel makes use of the paired-end read information to detect large deletions that fractured one of the paired reads (Figure 2.14). Although Pindel provides a reliable way to identify potential

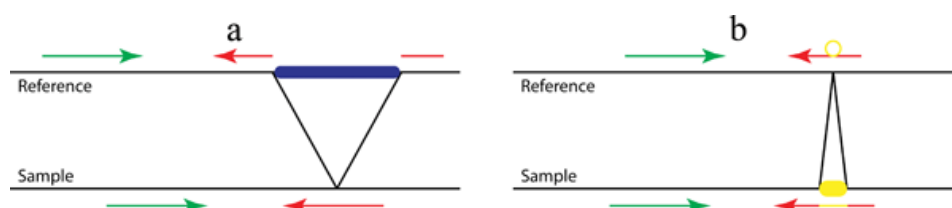


Figure 2.14: The pattern growth approach as used in Pindel. While a) depicts a large deletion that happened in the sample, b) indicates a smaller insertion. Using mapping information gained from both paired reads, Pindel determines paired reads that mapped with **Indels** or where only one end was mapped. This information is then used to identify potential variants Figure taken from (<http://gmt.genome.wustl.edu/packages/pindel/background.html>).

Indels, it lacks proper quality values. Therefore secondary values like read coverage and allele counts are used for variant filtration. Although this type of analysis greatly benefits from whole-genome sequencing data that, in principle, covers the complete genomic sequence, it can also be applied to panel or whole-exome sequencing data.

2.4.2 ClinVar, an archive for relationships among sequence variations and human phenotypes

When analysing a data set of a certain size, the pure number of individual **SNPs** quickly reaches a threshold at which it requires automatic processes to detect the most relevant variants, especially when considering the individual genetics of each patient. One way to tackle this challenge are public databases of human sequence variation. A commonly used source for this kind of information is ClinVar, a public archive of relationships among sequence variations and human phenotypes [105]. In its latest version (Aug 06, 2018), ClinVar contains a total of 700541 records, each providing a link between specific variants and human phenotypes.

To include this variant-phenotype associations into my analysis, I relied on the ClinVar VEP plugin that allows to automatically include ClinVar records during variant annotation. All variants were finally scanned for known associations to schizophrenia or related terms. This information was used independent of the previous analyses (2.3.2 - 2.3.5) to detect the presence of known schizophrenia-related variants.

2.4.3 The International Mouse Phenotyping Consortium

Even though the Human Genome Project was finished in 2003 and the initial sequence of the human genome was published in 2001 [106], the function of many of genes products remains unknown. One way to gain insights into this field are gene knock-out experiments that allow to investigate the function and genetic links of specific target genes. By using model organisms such as mouse or rat new knowledge can be inferred that helps to uncover gene functions.

One initiative that tries to systematically analyse each of the roughly 20,000 genes that make up the mouse genome is the International Mouse Phenotyping Consortium (IMPC) [107]. The IMPC is a global effort to identify the function of every protein-coding gene in the mouse genome. It currently contains 5,861 phenotyped genes that were phenotyped in 20 IMPC phenotyping centres, in 11 countries (Data Release 10.0). Data is generated by standardised allele production and is not aggregated from publications.

Using this data, I obtained significant experimental validated gene-phenotype associations for mental health related gene sets. The set of genes was collected using five available phenotype categories that include 1565 unique genes (abnormal brain morphology (n=72), abnormal fore-brain development (n=23), abnormal nervous system physiology (n=180), abnormal synaptic transmission (n=170) and behavior (n=1450)).

2.5 The Konstanz information miner

Success of large-scale data analysis depends on sophisticated bioinformatic support to process, integrate, analyse, and interpret Big Data volumes. In the following, I will briefly describe the open source platform Konstanz Information Miner (KNIME) [108] which I used as a processing interface for most of the presented analysis steps. I will also introduce KNIME4NGS [68], an NGS extension for KNIME which I developed in collaboration with several co-workers during the course of my thesis.

KNIME offers an intuitive graphical user interface, is user friendly and easily extendable. With a very strong and active community, a remarkable number of new functions have been incorporated into KNIME. As shown in Figure 2.15, KNIME allows to build and manage analysis workflows consisting out of multiple processing nodes. Each node can perform various tasks and can be connected into workflows. For NGS analysis, extensions like Knime4Bio [109] have been developed for the interpretation of biological NGS datasets starting from variant calls. However, a comprehensive toolbox of NGS nodes has been missing so far. Therefore, together with several colleagues (See [68]), I developed the KNIME4NGS, a comprehensive toolbox for next generation sequencing analysis [68]. In the following, I will present the main functionalities of KNIME4NGS which I used during the analysis of most datasets.

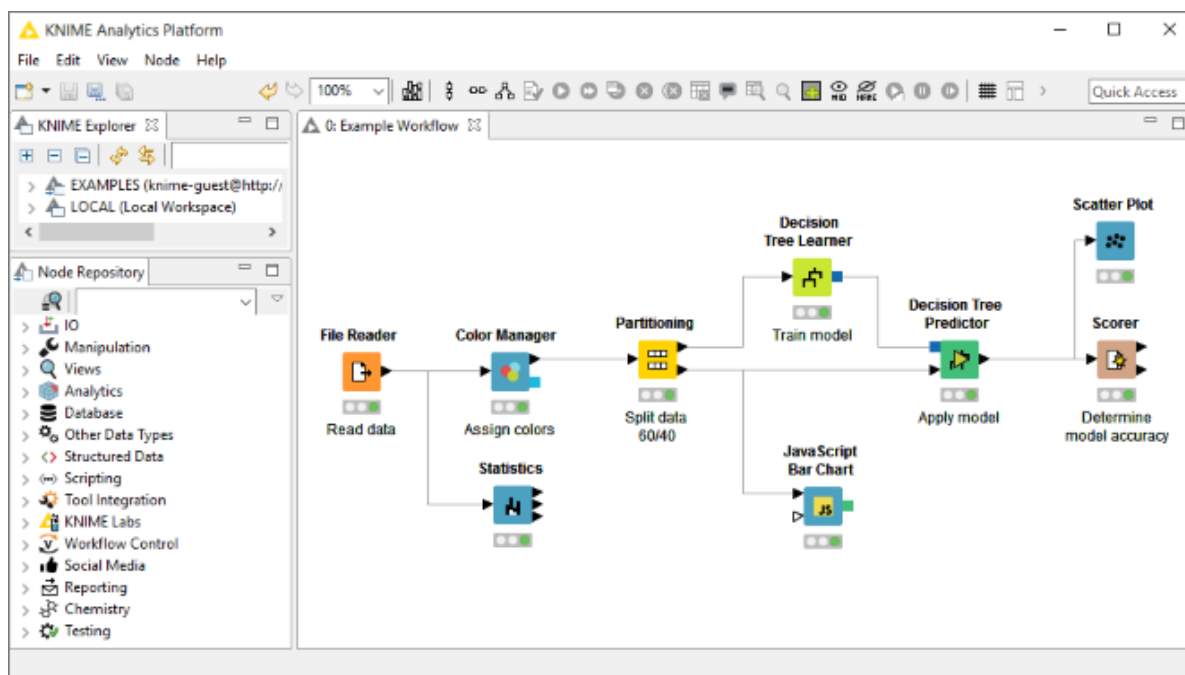


Figure 2.15: User interface of the Konstanz information miner (KNIME)

2.5.1 KNIME4NGS

KNIME4NGS [68] is an extension of [KNIME](#) that adds the functionality for essential [NGS](#) data processing. We developed a comprehensive linux-based [KNIME](#) toolkit including well-documented modules (nodes) for important steps like read preprocessing, read mapping, variant calling, detection of differential expression and annotation. Complementary to previously existing nodes, the toolbox now facilitates the assembly of basic building blocks into a wide range of customized [NGS](#) analysis workflows.

The official released KNIME4NGS package currently contains 42 nodes. Besides a number of auxiliary nodes, this extension provides building blocks and wrappers for well-established tools like BWA [69], DESeq [110] and [GATK](#) [75]. The nodes can be easily combined to create standardized workflows starting from initial preprocessing of raw data, up to the variant calling and biological annotation. This enables expert and non-expert users to set up reliable processing workflows quickly without the need of dealing with command-line tools. I also extended [KNIME](#) to use [NGS](#) specific file types. This improves workflow robustness enabling automated checks of tool configuration parameters and versions. To organise the required software binaries, I developed a lightweight binary management system accessible via the internal preference page. Thus, distributed nodes do not need to include the underlying executables, which simplifies keeping track with frequently updated software packages. The binary manager obtains necessary executables and integrates them into the workflow, minimizing repetitive node configuration. [NGS](#) datasets are typically very large. Parallelising workflows as much as controlling data

and results is crucial. Therefore we designed and tested our toolkit to work with corresponding [KNIME](#) extensions like openBIS [111], for storing and managing datasets, or the [KNIME](#) parallel chunk environment and cluster execution, for high-performance computing.

High-throughput Executor

In the analyses of large datasets, even established tools are prone to spurious premature termination. This aborts the entire workflow and requires extensive human intervention to ensure that all of the parallel executing nodes/samples finish successfully. To reduce this effort as much as possible, we developed the High-throughput Executor (HTE) as extension of the standard [KNIME](#) NodeModel. The HTE model collects process termination data of our nodes (e.g. error logs, execution time) and stores them in a local database provided by the system. Depending on the configuration and the process termination status, it automatically retries the execution of the failed node. The database allows the user to keep track and reduce the effects of unexpected software behaviour caused by for instance insufficient memory or randomly occurring errors. Additionally, the HTE model ensures by dedicated lock-files that not the entire workflow, but only those nodes are re-executed that depend on failed previous steps [68] .

RESULTS

The m4-Schizophrenia sequencing project contains exomes of 248 individuals with sporadic schizophrenia and 37 of their non-affected parents that were recruited from the Munich area in Germany. In the course of this thesis, I performed multiple analyses covering various aspects of the mutational spectrum. I applied an **NGS** analysis pipeline starting from initial quality control up to variant calling and annotation to gain new insights into the mechanism and neurophysiology of schizophrenia. In the following sections, I will present the most important results that were generated using the previously introduced datasets (Section 2.1).

3.1 Data quality, filtering and mapping

Validating the initial data quality is an inevitable step to avoid systematic and technical artefacts. In the following, I will present a compact overview about data quality during multiple pipeline stages which includes sequence read quality, target coverage, alignment processing and variant calling.

3.1.1 Read filtering

FastQC results of raw and filtered reads (Figures 3.1 and 3.2) showed a high quality of the generated sequencing data. Although the typical quality pattern for Illumina reads exist, which is indicated by quality decrease at both read ends, the results indicated no general sequencing problems. Besides the differing read lengths, there was also no detectable quality difference between data sequenced by Eurofins Medigenomix GmbH and the Technical University (WZW). The impact of read trimming is also clearly visible as the filtered reads showed almost no quality decrease towards read ends. The appearance of low quality bases at index position 1 for data

sequenced by Eurofins can be explained by low quality reads that were trimmed to a total read length of one nucleotide. The reads were not removed completely as this would result in a loss of order within the respective sequence read files whenever the paired sequence is not discarded. In addition to base quality, I also investigated the remaining quality scores that are provided by the FastQC method. Among others, this includes the number of bases that were called as 'N', which indicates that no clear evidence for any of the four nucleotides was found, or the overall GC-content. None of these values indicated general quality issues regarding the whole dataset.

3.1.2 Read mapping and target region coverage

Filtered reads were mapped against the UCSC human reference genome (build hg19). The mapping was done using the Burrows-Wheeler Aligner (BWA 0.7.12-r1039). Overall read and mapping statistics are shown in Figure 3.3. On average, a sample was covered by about 83 million reads. For most individuals, almost 100% of the reads could successfully be assigned to a genomic position with only a few samples showing numbers below 95%. Of all reads, about 63% were mapped into to primary target region (SeqCap EZ Exome v3). Here, two sample clusters are visible; while for samples sequenced by Eurofins, about 75% of the reads fall into the target region, samples coming from the TUM Weihenstephan show lower numbers (~ 50%). This observation can be easily explained as TU samples were sequenced by using a slightly different capture kit, therefore partially include different genomic regions. On average, about 52 million reads were mapped onto the SeqCap target region.

3.1.3 Alignment processing, quality filtering and variant calling

Finalized alignments, were further processed (See Section 2.2.1) to improve mapping quality. Furthermore, I removed four trios due to overall quality concerns, based on previous *de novo* variant analyses. All four trios showed unusual *de novo* patterns, clearly diverging from other trios. Overall, the *de novo* rate was increased ~10 fold compared to the remaining samples, raising questions about the reliability of these four trios. In addition, another trio was discarded due to unclear family structure. All remaining trio samples as well as all individual samples were used as input for the following variant calling. I jointly called all samples, resulting in 672,550 raw SNPs and Indels. After applying GATK's VariantQualityScoreRecalibration and further filters (See Section 2.2.3), the final call set of 218,500 variants remained.

3.2 Analysis of genetic variation types

After finalizing the variant call set, I performed several analyses each focusing on specific variant types or annotation features (*de novo* variants, compound effects, edgetics) to gain a comprehensive set of candidate genes. In the following, I will present the most relevant results.

3.2. ANALYSIS OF GENETIC VARIATION TYPES

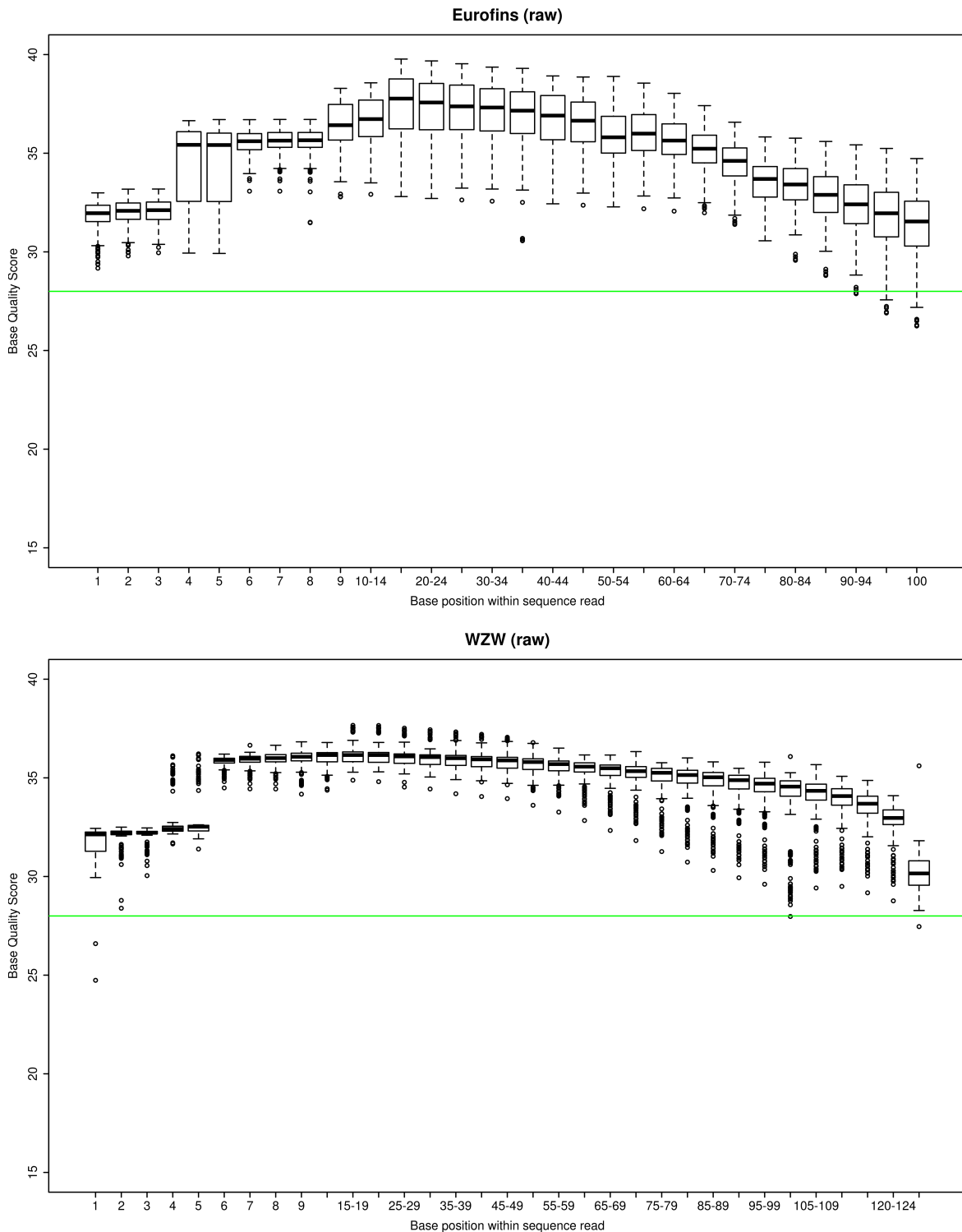


Figure 3.1: Quality profile showing the base quality per position index for raw sequence reads, separately for sequence data from Eurofins Medigenomix GmbH and the Technical University (WZW). In both cases, raw quality scores indicate an already high overall quality. However, as expected, quality scores at the beginning as well as the end of the read drop towards or below the high quality threshold as indicated by the horizontal line (green).

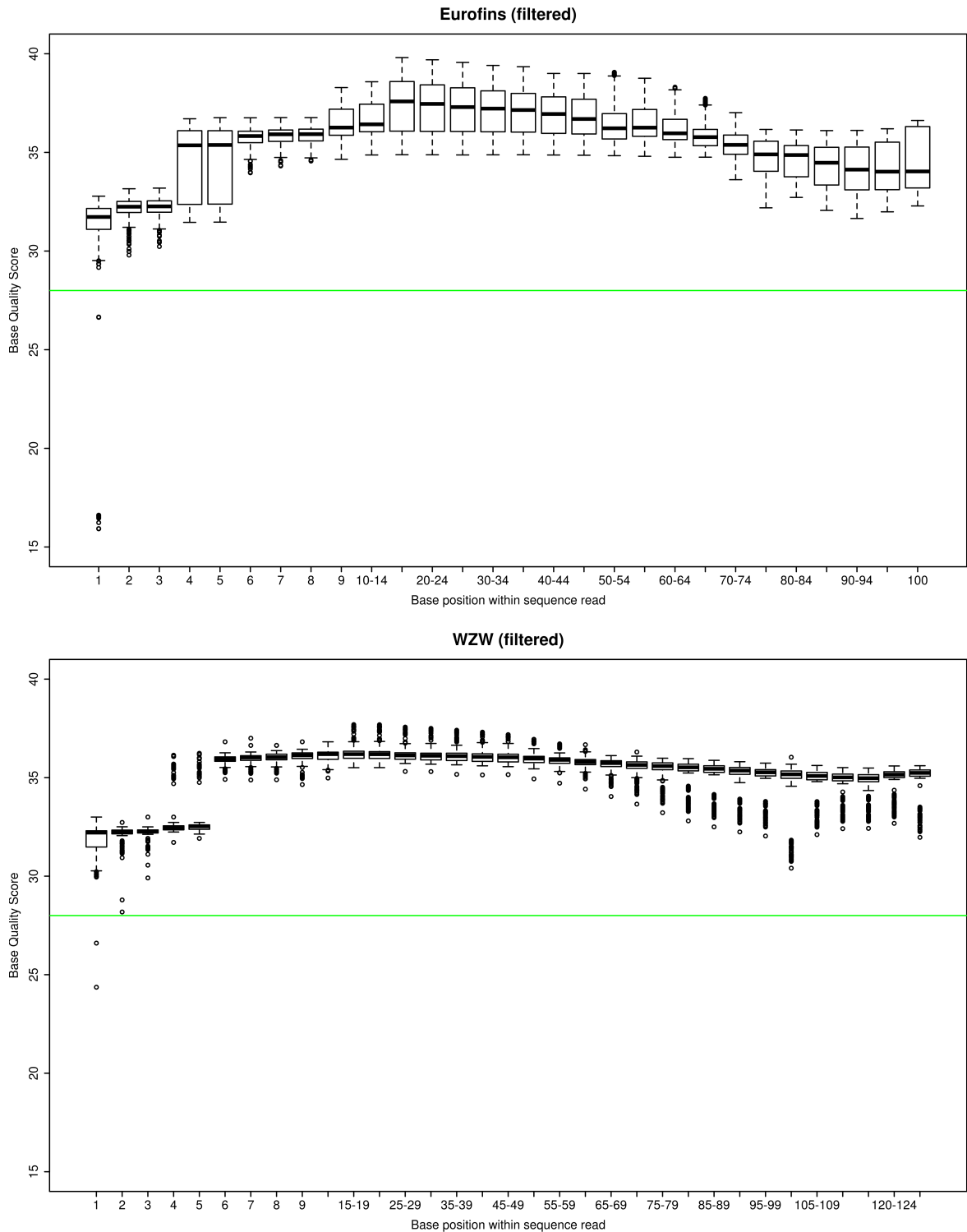


Figure 3.2: Quality profile showing the base quality per position index for filtered sequence reads, separately for sequence data from Eurofins Medigenomix GmbH and the Technical University (WZW). In comparison to the raw quality profile, the overall quality of the remaining bases/reads has increased especially towards the read ends.

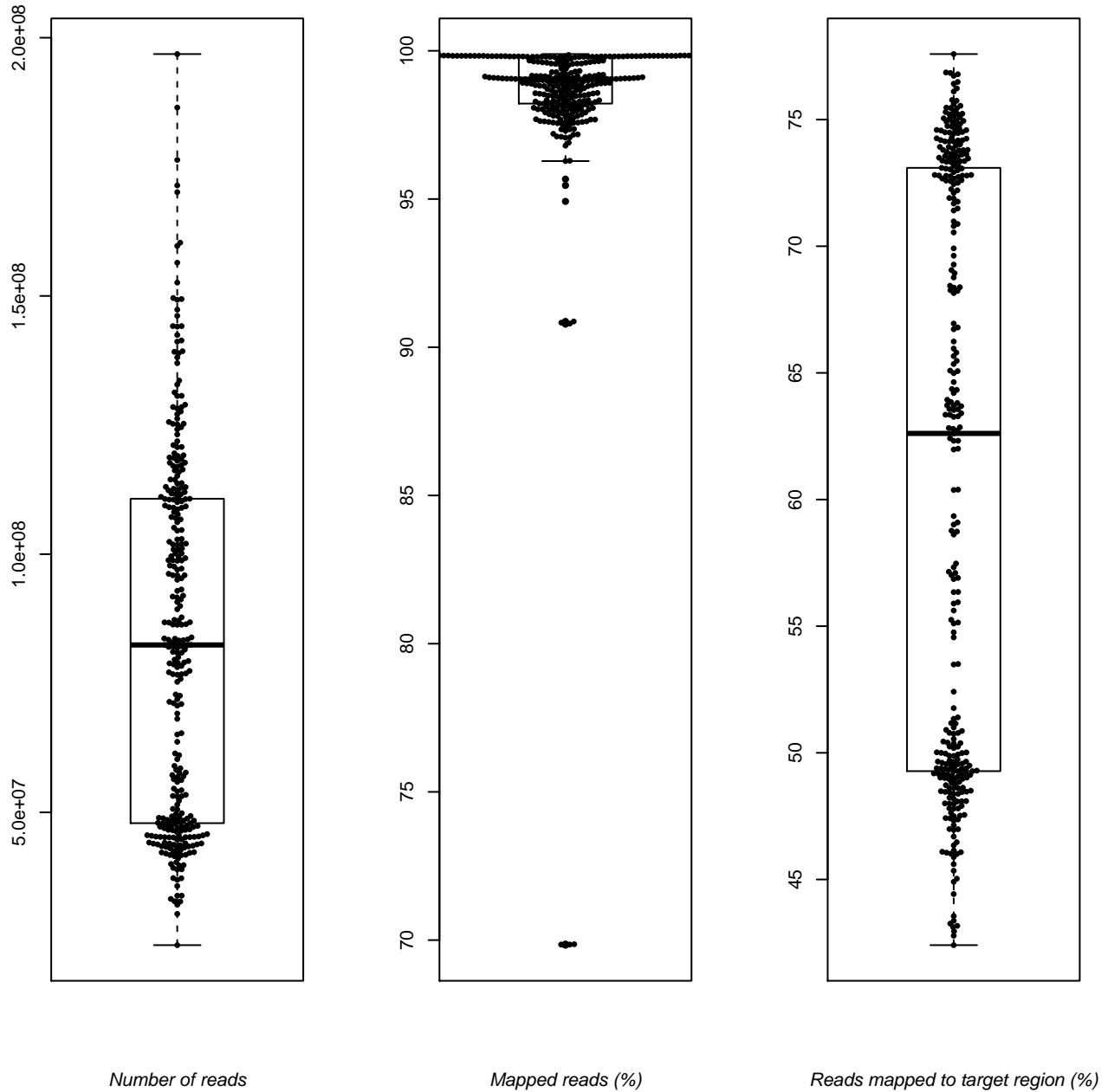


Figure 3.3: Beeswarm plot showing overall mapping statistics for all included samples. The left panel indicates the total number of sequence reads that were generated for each sample. The second panel presents the percentage of reads that were successfully mapped to the human reference genome (hg19). On average this was true for about 99% of the reads which equals a total amount of 81 million sequence reads. The right panel presents the percentage of reads that were mapped to primary target region (SeqCap EZ Exome v3). The two clusters that are visible in this panel can be explained by the different capture kit that was used for samples sequenced at TUM Weihenstephan.

3.2.1 *De novo* variants

As shown by large sequencing studies [56], *de novo* variants strongly contribute to schizophrenia. Under the filter criteria applied, I detected 74 potential exonic *de novo* variants (Figure 3.4) with a nonsynonymous to synonymous mutation rate of 3.13. I used several external databases (1000 Genomes Project, dbSNP, Exome Sequencing Project (ESP) and ExAC) and the CADD deleteriousness score to extract a high confidence set of rare deleterious variants, leaving 46 (36 missense, 5 frameshift indels, 3 stopgain, 2 splice site) putative *de novo* variants. These 46 mutations distribute among 24 of the 32 families (Figure 3.5).

I did not find any evidence for a correlation between the number of *de novo* variants and the age of the respective father. Still, this observation might be due to the small sample size, therefore a replication using a larger sample set could provide further insights.

The overall *de novo* mutation rate is 1.8×10^{-8} with an 95% confidence interval of $[1.49 \times 10^{-8} - 2.38 \times 10^{-8}]$ ($p=1.61 \times 10^{-5}$). This is significantly different to the expected neutral rate of 1.1×10^{-8} . This observation is in accordance to the reported increased frequency of *de novo* variants in schizophrenia [49, 51].

Of the 46 candidate mutations (Table 3.1), three were located in a genetic locus, previously associated with schizophrenia [112]. Furthermore, 13 genes have a direct protein-level interaction to at least one of the genes located within the 108 schizophrenia loci detected by the Schizophrenia Working Group of the Psychiatric Genomics Consortium.

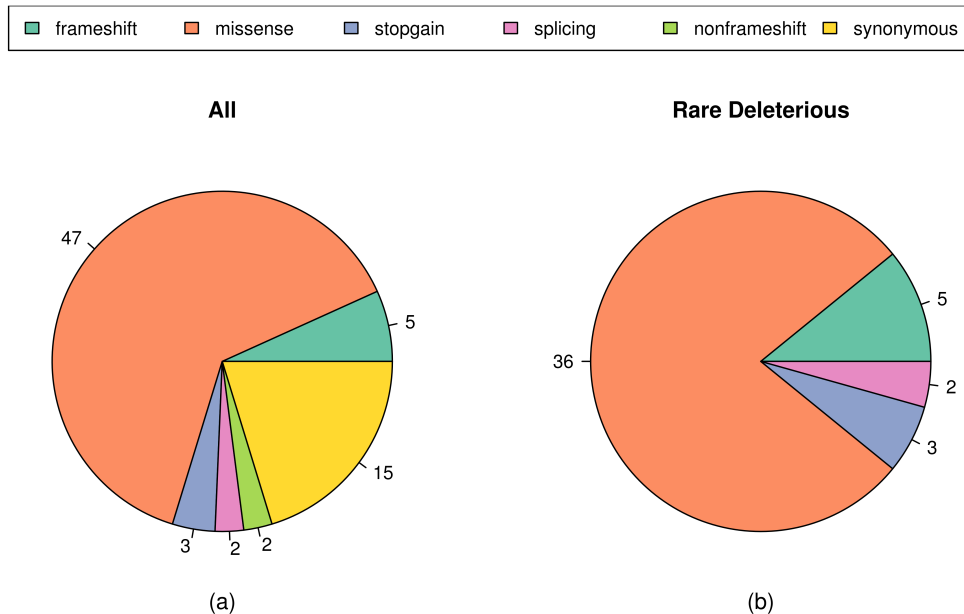


Figure 3.4: Overall number of *de novo* mutations, grouped according to their genomic impact. Part (a) includes all 74 *de novo* variants, (b) the 46 rare deleterious *de novo* variants.

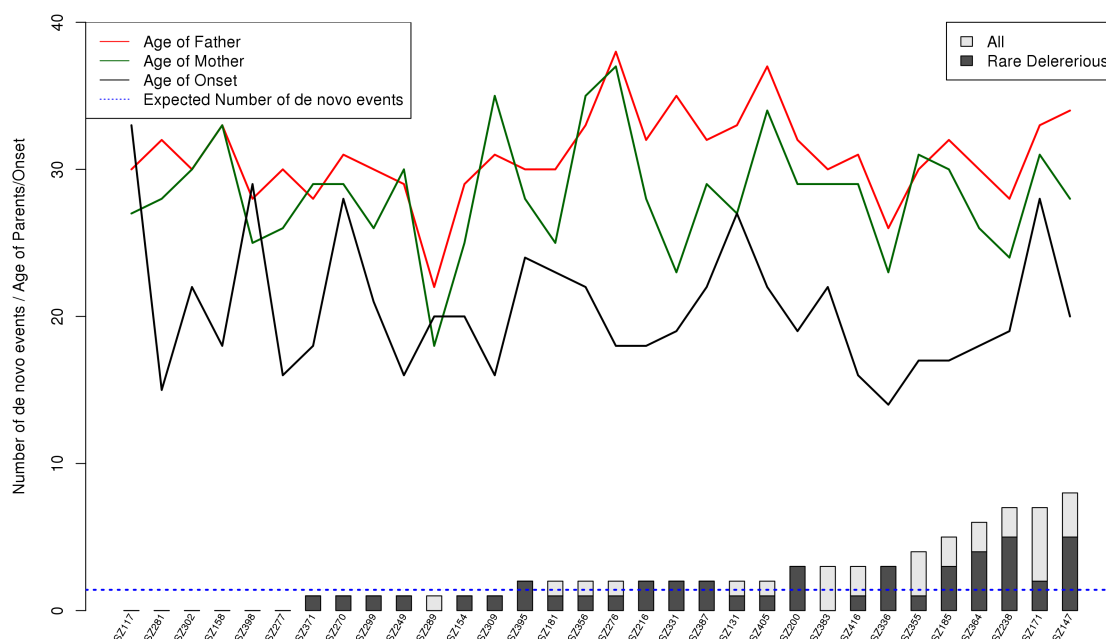
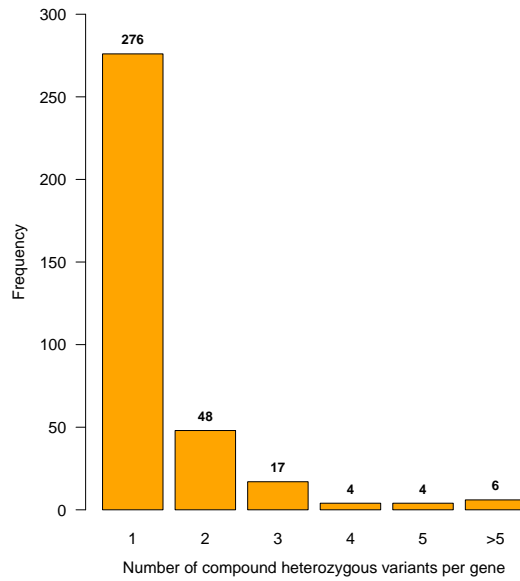


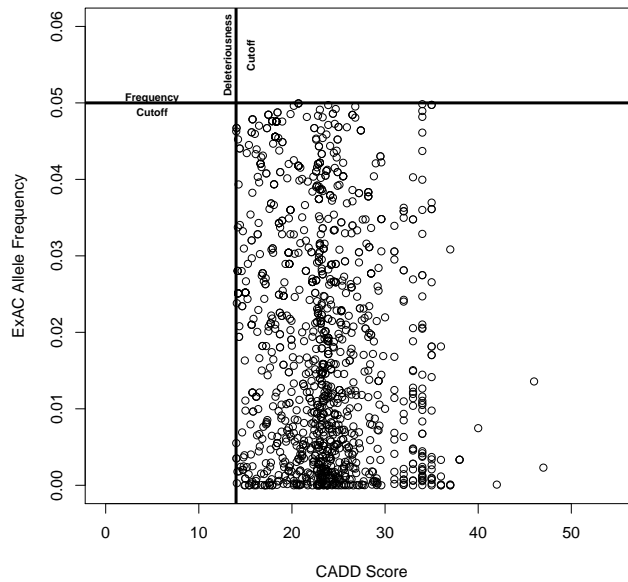
Figure 3.5: Number of *de novo* variants for each affected trio offspring. Each bar represents the number of all and filtered (rare deleterious) *de novo* variants whereas the blue dotted line indicates the number of expected variants based on the neutral *de novo* mutation rate of 1.1×10^{-8} . Additionally, the age of both parents (green and red line) as well as the age of onset (black line) are shown.

3.2.2 Compound heterozygous variants

Besides focusing on single *de novo* events with potential disease-relevant effects, I also investigated the presence of rare compound heterozygous variants. This offers the possibility to reach beyond the prediction of effects of single variants and instead takes into account that complex diseases are often caused by multiple genes and/or multiple mutations on individual genes [86–90]. Compound heterozygous variants can have similar effects like homozygous **LoF** variants as long as they affect both DNA alleles. Again, I relied on ExAC and CADD to extract a list of rare deleterious mutations and detected 354 genes carrying at least one compound heterozygous variant (Figure 3.6). 79 genes were affected in multiple samples whereas six genes were found more than five times. As expected, with decreasing allele frequencies, the annotated CADD scores tend to increase as deleterious variants that affect natural fitness should be removed by purifying selection. Overall, I did find more compound heterozygous variants in trio samples than in individuals without parental data (Figure 3.7). This effect can be explained as there is less phasing information for non-trio samples available which limits the number of potential variants.



(a)



(b)

Figure 3.6: Compound heterozygous variants in schizophrenia. **(a)** indicates the number of genes (coding region) affected by one or more compound heterozygous variants. Six genes were affected more than five times which, however, is most likely explained by the median protein sequence length of 6777 that is clearly above average (Median length of human proteins classified in Pfam-A: 416 [113]). **(b)** shows the distribution of the respective ExAC allele frequencies and CADD scores that are slightly negative correlated (-0.15 with a 95% confidence interval [-0.21,-0.10] and p-value 3.96×10^{-8})

3.2. ANALYSIS OF GENETIC VARIATION TYPES

Chr.	Pos	Gene	Feature	Mutation	MaxDBAlAF	Candidate	CADD	ID
1	8390628	SLC45A1	missense	c.1177A>G;p.T393A		Y	24.4	SZ238
1	883543	NOC2L	missense	c.1627T>C;p.F543L		PPI	25.9	SZ364
1	110169447	AMPD2	missense	c.550A>C;p.T184P	0.002643	PPI	26.7	SZ336
1	222825612	MIA3	missense	c.4024G>A;p.V1342M			33	SZ336
2	74708611	CCDC142	missense	c.1076T>C;p.L359P			18.37	SZ364
2	37113948	STRN	frameshift	c.T953insA+;p.M318fs		PPI	31	SZ309
2	180008405	SESTD1	stopgain	c.763C>T;p.R255*			36	SZ171
2	196866551	DNAH7	stopgain	c.1021C>T;p.Q341*			37	SZ355
3	138023763	NME9	missense	c.560T>C;p.V187A			15.26	SZ147
3	108698515	MORC1	splicing	c.2325-1G>T			17.81	SZ364
3	108163500	MYH15	missense	c.2702T>C;p.L901P			28.3	SZ147
5	147845469	HTR4	missense	c.1096A>G;p.S366G			17.87	SZ216
5	31983313	PDZD2	missense	c.529C>T;p.R177C			34	SZ270
6	152651292	SYNE1	frameshift	c.G14312delG-;p.A4772			24.9	SZ131
6	56463354	DST	stopgain	c.3979C>T;p.Q1327*			38	SZ405
6	155500195	TIAM2	missense	c.2546A>C;p.H849P		PPI	22.1	SZ387
6	87953243	ZNF292	missense	c.792A>C;p.E264D			24.5	SZ146
6	122744801	HSF2	frameshift	c.A1147del-T;p.F383			35	SZ147
7	149523649	SSPO	missense	c.14560G>A;p.G4854S			16.36	SZ185
7	101833124	CUX1	missense	c.1082A>G;p.D361G		PPI	25.1	SZ356
7	97846708	TECPR1	frameshift	c.G3489delG-;p.C1164			26.5	SZ147
7	150920795	ABCF2	missense	c.691G>C;p.G231R			31	SZ249
8	141545637	AGO2	missense	c.2201C>T;p.T734M	0.000008132		33	SZ364
9	130863590	SLC25A25	missense	c.391C>T;p.R131C	0.0001626		22.9	SZ387
9	26984497	IFT74	missense	c.405G>T;p.R135S			26.2	SZ395
10	97626013	ENTPD1	missense	c.1427C>T;p.S476F		PPI	23.5	SZ238
10	1151131	WDR37	missense	c.1027C>T;p.R343C	0.000008132		34	SZ371
11	124873761	CCDC15	splicing	c.2215-2A>G			22.6	SZ276
12	57436930	MYO1A	missense	c.1024C>T;p.R342W	0.0001952	Y	35	SZ154
12	124288286	DNAH10	missense	c.2339C>T;p.A780V	0.00001626		22.4	SZ331
13	45827101	GTF2F2	missense	c.482A>G;p.Y161C	0.000008133	PPI	19.95	SZ185
14	29237038	FOXG1	missense	c.553A>G;p.S185G		PPI	25.2	SZ238
14	68048877	PLEKHH1	missense	c.3376T>G;p.F1126V	0.049		32	SZ238
16	88873828	CDT1	missense	c.1415C>T;p.A472V	0.000459		16.16	SZ336
16	29974526	TMEM219	missense	c.62T>C;p.V21A		Y	22.2	SZ331
17	40729317	PSMC3IP	missense	c.139G>T;p.V47L			33	SZ171
17	7915774	GUCY2D	missense	c.1963A>G;p.R655G		PPI	23.5	SZ185
17	38563812	TOP2A	missense	c.1615A>G;p.M539V		PPI	26.5	SZ299
17	42850632	ADAM11	missense	c.829T>G;p.Y277D			24.8	SZ181
18	57122114	CCBE1	missense	c.623T>C;p.M208T	0.000008132		23.3	SZ200
18	5891500	TMEM200C	missense	c.563A>G;p.H188R	0.000008158		20.7	SZ395
19	15562682	RASAL3	missense	c.2960T>C;p.L987P			15.7	SZ200
19	54666439	TMC4	missense	c.1469T>C;p.L490S			26.7	SZ147
19	3977260	EEF2	missense	c.2336C>T;p.T779I		PPI	27.4	SZ216
20	47351125	PREX1	frameshift	c.G477insT+;p.L159fs		PPI	35	SZ200
X	51637831	MAGED1	missense	c.154C>T;p.R52C	0.000008172	PPI	23	SZ238

Table 3.1: The 46 candidate *de novo* mutations found in 32 trios. For each variant, several annotations are provided. This includes the maximal allele frequency (MaxDBAlAF) across the used databases(1k201304, dbsnp138, ESP6500AA, ESP6500EA and ExAC), an indication whether the gene is located in a schizophrenia candidate (Y) locus or has an interaction with it (PPI), the CADD score as well as the affected individual.

3.2.3 Loss-of-Function variants

LoF variants are the most severe form of a **SNP** or short **Indels**. Instead of leading to protein-level aberrations, **LoF** variants lead to a complete loss of the gene product, given a homozygous genotype that affects all transcripts. Here I investigated the overall number of **LoF** mutations (independent of the actual genotype) and used phase information to detect genes that were knocked out in affected offsprings but remained functional in both parents.

In total, I found that the analysed exomes of schizophrenia patients contain 139.17 ± 11.45 (mean \pm sd) putative **LoF** variants (Figure 3.8). 7.81 ± 2.41 genes are completely inactivated. When filtering for rare variants, 63.18 ± 9.64 **LoF** mutations remain with 2.59 ± 1.20 genes being affected by a full knock-out. There was no significant difference compared to healthy parents as they

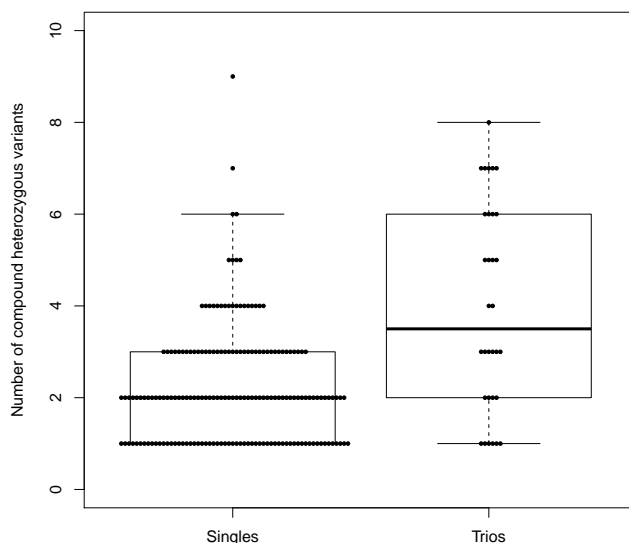


Figure 3.7: Number of rare deleterious compound heterozygous variants for trio and non-trio data. Overall, more variants were detected in trio data than in non-trio samples. This can be explained by the reduced phasing information that can be generated for individual samples.

showed almost similar numbers (LoF: 141.24 ± 12.24 , Knockout: 7.83 ± 2.03) (For a comparison against current literature, see Section 4.1.3.)

Additionally, I also looked for rare LoF variants that affect all isoforms of the respective gene. Such compound LoF mutations may either be due to homozygous LoF variants or a combination of multiple LoF variants. Such events lead to a complete knock-out of the gene product and therefore have severe impact on all related downstream processes. Here I found 10 genes that were knocked out in cases but remain functional in both respective parents and were therefore added to the list of candidate genes.

3.2.4 Genetic Burden

A third compound effect that was investigated describes the enrichment of inherited/*de novo* deleterious variants. For each family, I calculated the genetic burden (See Section 2.3.4), to detect genes that accumulate deleterious variants. I kept all genes that showed an increased genetic burden in cases compared to their parents in more than 10 samples. As shown in Figure 3.9, of 27192 canonical genes defined by UCSC, only 7644 have an increased burden in at least one patient. This number quickly drops when increasing the cutoff of affected patients leaving 35 candidate genes that matched the specified frequency criterion.

An exemplary depiction of one those 35 candidates is shown in Figure 3.10. Here, samples can be grouped into three main categories. Whereas some patients show a clear increase of the overall genetic burden, others show no difference or a decrease of deleterious variants. Taking

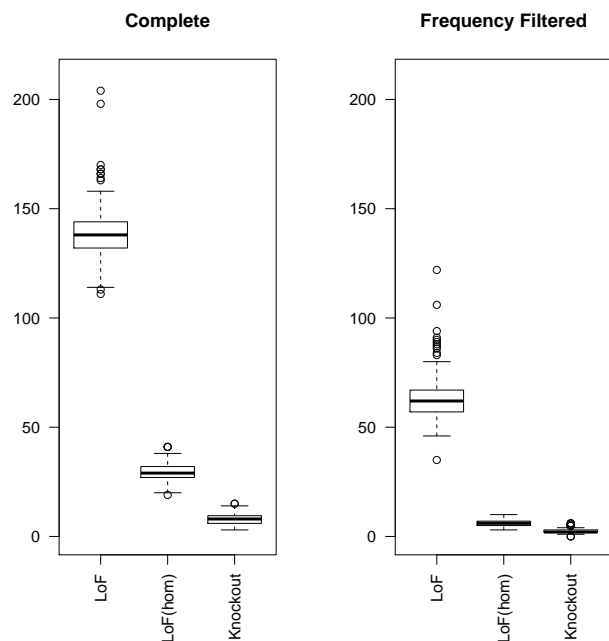


Figure 3.8: Number of **LoF** variants for the complete and frequency-filtered variant set for schizophrenia patients. In both cases, the overall number of **LoF** variants, homozygous **LoF** variants and mutations that affect all transcripts (at the respective site) and lead to complete knockout of the gene product are given.

those observations into account, it becomes clear that an individual typically does not harbour deleterious variants. Still, certain recombination or *de novo* events can lead to changes towards an increase of the genetic burden, making those genes a potential target for further investigations.

3.2.5 Edgetics

After focusing on single *de novo* events as well as compound effects, I included a further layer of information by using the dSysMap database of protein-protein interaction (PPI) interfaces. The database contains specific amino acid sites that interact during PPIs and therefore allows to map mutations on protein structures and the human interactome. In this case, not only the mutated protein is affected, but also the PPI-partner throughout the potential loss of interaction. I scanned all variants for such interaction-disrupting mutations and found 31 variants that were located within a PPI interface. This includes 20 homomeric PPIs (Interactions between the same type of protein) and 11 heteromeric PPIs (Table 3.2). For six proteins, I found multiple interaction-disrupting mutations. In total, 37 unique genes were found to be affected (28 mutated proteins and 9 additional interaction partner).

One of the 11 heteromeric PPIs, the interaction between ACY3 (Aminoacylase 3) and ASPA (Aspar-toacylase) which was found to be affected by a non-synonymous mutation (c.C541A;p.P181T), is shown in Figure 3.11. Here, the cyclic and nonpolar proline is exchanged by a hydroxyl-containing

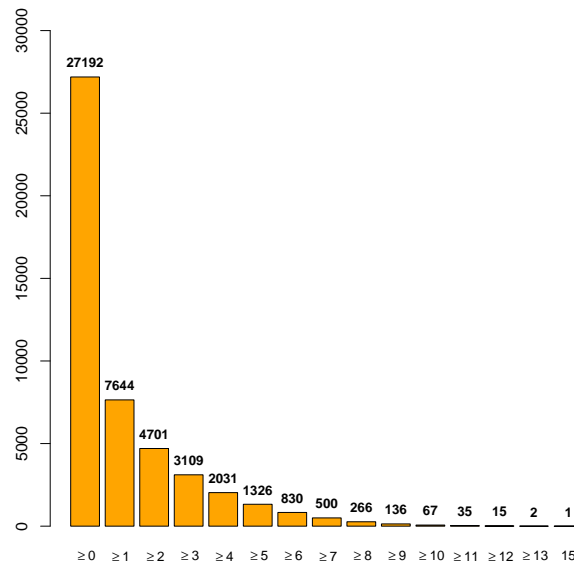


Figure 3.9: Overview about the increase of genetic burden for the 27192 canonical genes defined by UCSC. Each bar represents the number of genes for which at least n offspring samples have an increased genetic burden.

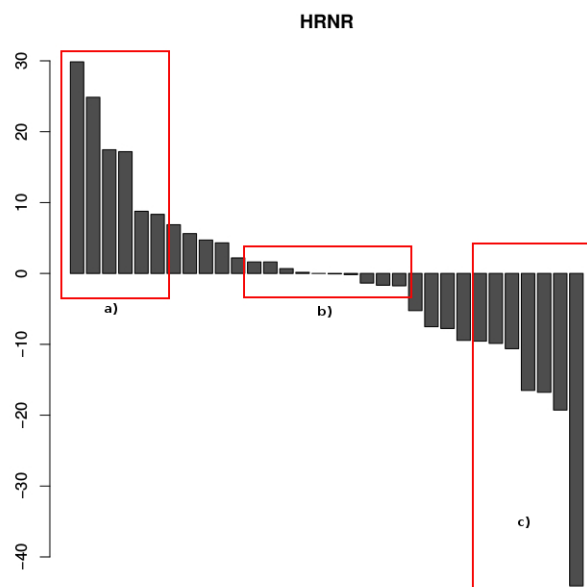


Figure 3.10: Exemplary genetic burden difference (CSS_{Gene} difference) between affected offspring and the respective parents for Hornerin (HRNR), a gene involved in calcium ion binding. a) indicates individuals with increased burden, b) with equal burden and c) with decreased burden.

3.2. ANALYSIS OF GENETIC VARIATION TYPES

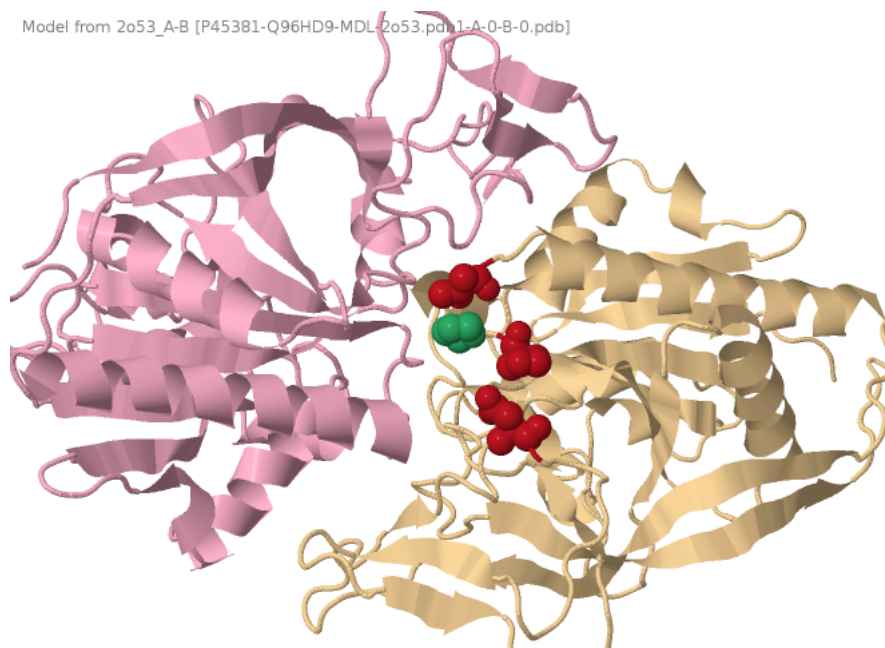


Figure 3.11: Interaction between ACY3 (Aminoacylase 3) and ASPA (Aspartoacylase). The mutated residue is shown in green, the remaining PPI interface residues in red. Here, the mutated Proline at position 181 of the Aspartoacylase protein affects the direct interaction with Aminoacylase 3.

and polar threonine. Although the overall molecular mass is similar, the change of biochemical properties clearly has the potential to critically disrupt the interaction of both proteins.

Chr.	Genomic Pos.	Mutation	Mutated Gene	Interaction Partner	ExAC Frequency	Samples
chr1	183532364	c.A1256T:p.N419I	NCF2	NCF4	0.005904	SZ129, SZ143, SZ223, SZ297, SZ364, SZ156
chr3	52485426	c.C435A:p.D145E	TNNC1	TNNT2	0.0004086	SZ221, SZ117, SZ319
chr9	37783990	c.A395C:p.D132A	EXOSC3	EXOSC9	0.0005095	SZ148
chr10	43613908	c.A2372T:p.Y791F	RET	EGFR	0.002644	SZ353, SZ249, SZ342
chr11	111779556	c.G460A:p.G154S	CRYAB	HSPB2	0.001274	SZ104
chr13	37583420	c.G815C:p.S272T	EXOSC8	EXOSC9	0.005861	SZ377, SZ236, SZ416, SZ138
chr13	51519581	c.G529A:p.A177T	RNASEH2B	RNASEH2C	0.001925	SZ365, SZ270, SZ240
chr15	72105906	c.C925T:p.R309W	NR2E3	PPARG	0	SZ146
chr17	3392543	c.C541A:p.P181T	ASPA	ACY3	NA	SZ412
chr19	35775902	c.G212A:p.G71D	HAMP	LCN2	0.003521	SZ271, SZ241, SZ375, SZ243, SZ363, SZ316, SZ155
chr19	12924251	c.C871T:p.R291C	RNASEH2A	RNASEH2C	0	SZ212

Table 3.2: Heteromeric protein interactions affected by rare interaction-disrupting mutations. For each interaction, the involved proteins, the ExAC frequency of the variant as well as the affected samples are given.

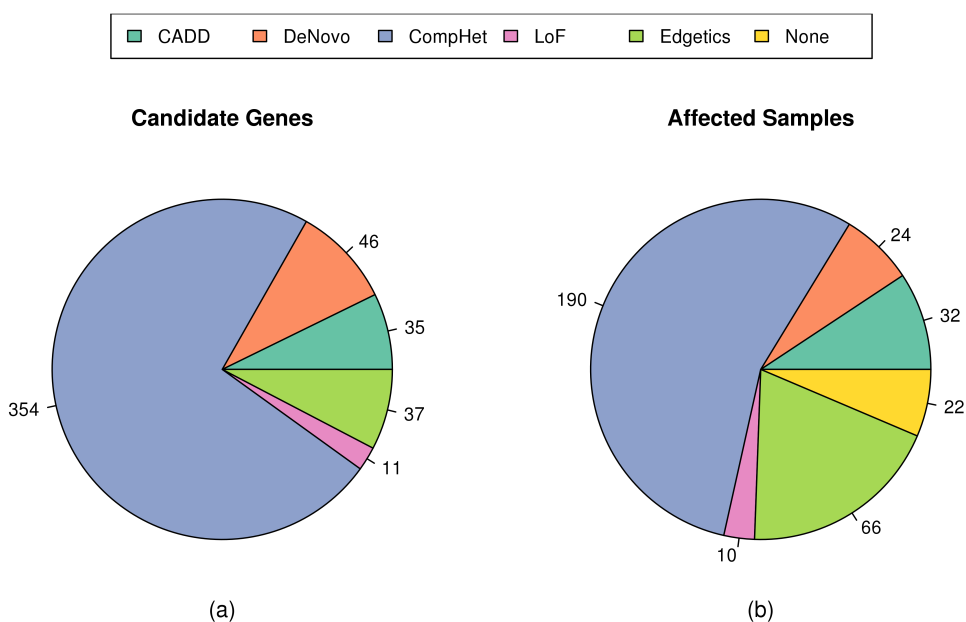


Figure 3.12: Number of candidate genes resulting from the different analysis methods (a) and the respective number of affected samples (b). As samples are partially affected by multiple variant types, there exist sample overlaps between certain affection groups. For individuals grouped into "None", I did not find any potential disease-relevant variants.

3.3 Candidate genes and integrative analysis

I have performed several analysis steps to create a comprehensive list of potential disease-contributing genes. This resulted in a total of 468 candidate genes (Figure 3.12) that are affected by different types of mutations, each covering certain aspects of the genetic mutation spectrum. 14 genes (*AHNAK2*, *AKAP13*, *AQP7*, *CCDC60*, *CDT1*, *CSMD2*, *DNAH10*, *DNAH8*, *MPO*, *MYO18B*, *NCF2*, *PDZD2*, *SYNE1*, *DNAH7*) were found by at least two of the five (Figure 2.1) approaches in at least a single sample. Overall, these candidate genes cover all but 22 individuals. In the following, I will present the results of the integrative candidate gene analysis that makes use of all potential disease-contributing genes. I used protein-protein interaction networks and general enrichment methods to find common characteristics and shared pathways among all candidates. Additionally, I checked all candidates for existing schizophrenia-related mouse phenotypes.

3.3.1 Protein-protein interaction network

As a first step of the general investigation of the obtained set of candidate genes, I created a comprehensive PPI interaction network to detect potential hub proteins as they may yield further insights about important pathways and interactions in the pathogenesis of schizophrenia. This was done by using all 468 genes and a first layer of direct interaction partners (Figure 3.13). The

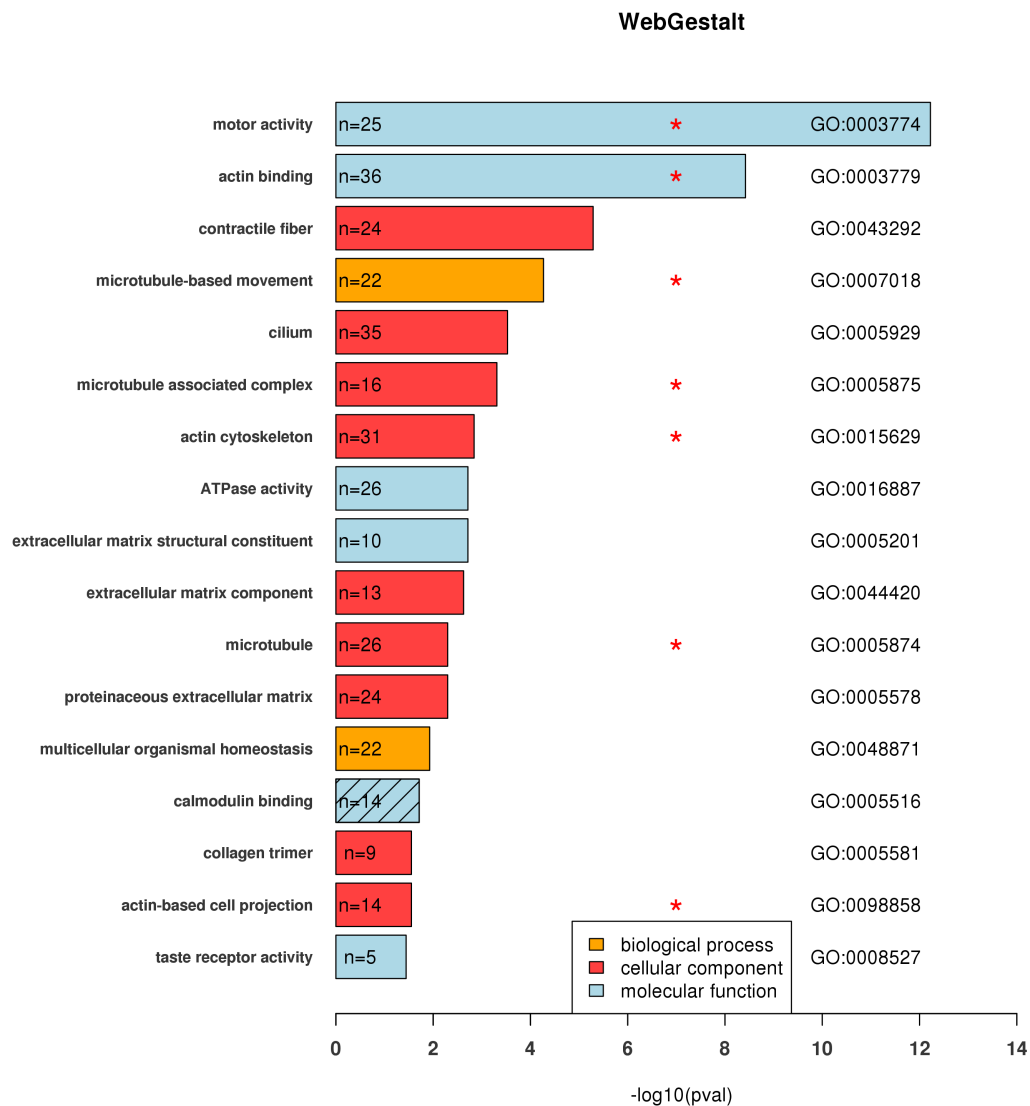


Figure 3.14: Enriched GO terms for the three GO categories (Molecular function, Cellular component, Biological process). Microtubule, actin and motor protein related terms are marked with a red *. For each category the respective GO ID and number of candidate genes falling into this category (n) are given. Hatched bar indicates that the group was discarded after DNENRICH analysis.

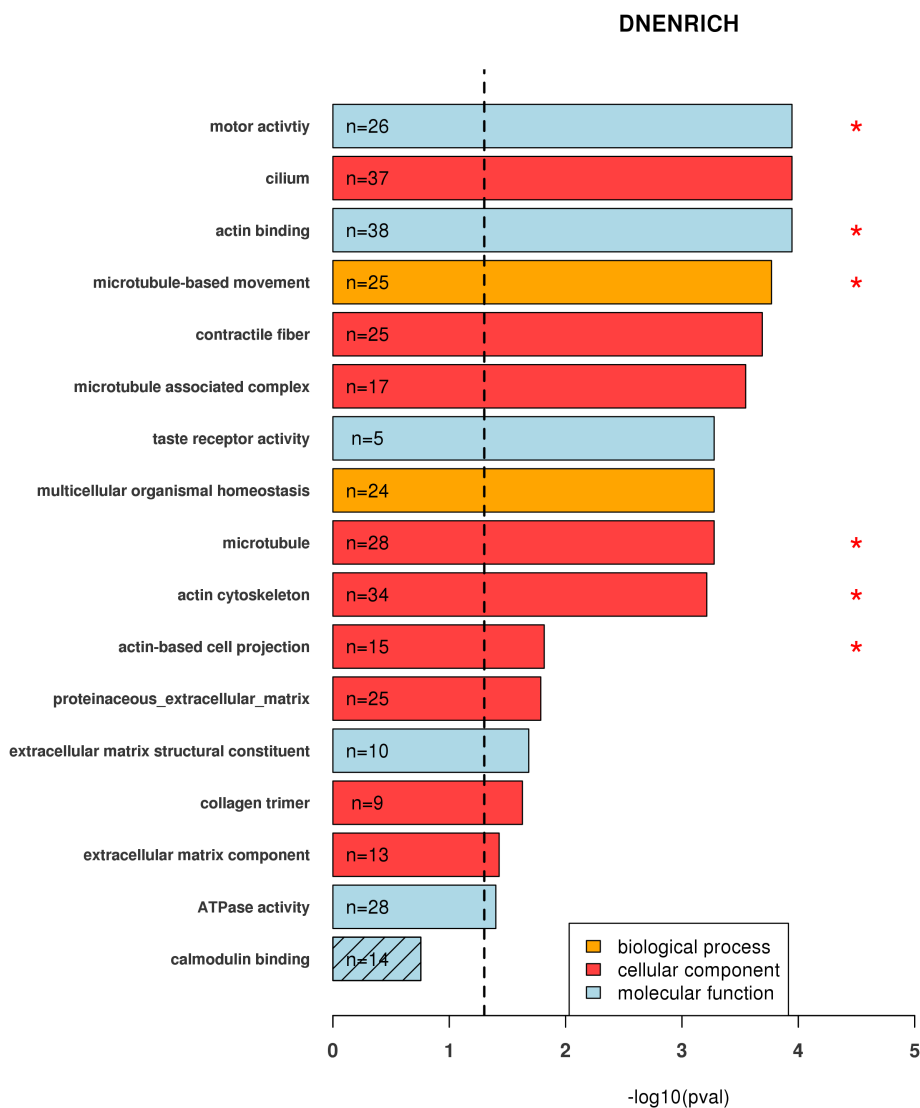


Figure 3.15: Enriched GO terms after DNENRICH analysis. Microtubule, actin and motor protein related terms are marked with a red *. For each category the number of candidate genes falling into this category (n) are given. Hatched bar indicates that the group not significantly enriched. The dashed line indicates the applied p-value cutoff.

To ensure that the enrichment results were not biased by gene length, I applied DNENRICH to perform a second enrichment analysis. As the original background GO sets of WebGestalt were not available, I retrieved them by using AmiGO, an online access to ontology and annotation data [114]. As a result, there exists a small variance in the number of genes that fall into specific categories.

Besides a single set (calmodulin binding), all tested categories (No new categories were included in the analysis) remained significant after DNENRICH analysis. A graphical representation of the results is shown in Figure 3.15.

The remaining top scoring categories clearly show an enrichment of microtubule and actin-related genes throughout all GO categories (Biological process, Molecular function, Cellular component) as seven of 16 significant groups are directly related to one of those components. For the microtubule-based movement and motor activity sets, this is mainly due to the frequent occurrence of disrupted motor proteins dynein, kinesin and myosin that occurred during various analysis steps. Furthermore, I found several affected centrosomal protein genes as well as a number of microtubule and actin-associated proteins that explain the enrichment of microtubule and actin cytoskeleton sets.

To investigate whether the enrichment is based on a selection of few samples or due to a larger part of the dataset, I generated sample-specific counts for affected genes of the neuronal cytoskeleton (microtubule, actin cytoskeleton, motor activity sets) as shown in Figure 3.16. Of the 288 schizophrenia patients, 185 had at least a single damaging variation type within a gene of the neuronal cytoskeleton. This accounts for almost 2/3 of all schizophrenia samples. Close to 1/3 even had multiple damaged genes with 8 being the highest detected number for a given sample. In contrast, for 103 patients, I did not find any evidence for disruptions in the neuronal cytoskeleton.

3.3.3 Genes with schizophrenia-related mouse phenotypes

Approximately one-third of all mammalian genes are essential for life [107]. Knockout experiments in mice have provided tremendous insight into gene function and related disorders. To further explore the obtained candidate list, I used data of the IMPC [107] to investigate the presence of mouse models with schizophrenia-related phenotypes.

Candidate genes were compared against five IMPC gene categories that include a total of 1565 unique genes (abnormal brain morphology (n=72), abnormal forebrain development (n=23), abnormal nervous system physiology (n=180), abnormal synaptic transmission (n=170) and behavior (n=1450)).

This resulted in an overlap of 36 genes with significant mouse phenotype in at least one the five mentioned categories. Amongst others, these phenotypes include observations like general abnormal behavior, decreased grip strength as well as abnormal brain morphology and decreased prepulse inhibition. For 32 of 36 genes a link of the respective phenotype to schizophrenia was

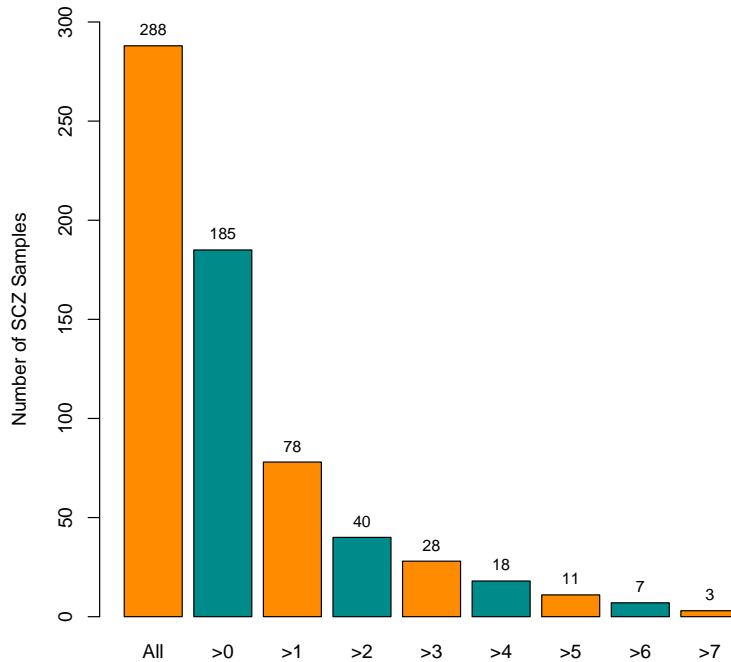


Figure 3.16: Damaging genetic variants in genes of the neuronal cytoskeleton. For the 288 samples with diagnosed schizophrenia, the number of patients with a certain amount of mutated genes of the neuronal cytoskeleton is given for different cutoffs (>0 - >7).

found using current literature (Table 3.3).

3.4 Analysis of structural variants and association with known candidates

Since schizophrenia is complex disease which has yet to be fully explained, several hypotheses about the genetic causes exist. This involves a large number of putative candidate genes as well as CNV and structural variants like the 22q11 deletion. Therefore, I investigated the presence of large **Indels** on chromosome 22 as well as variants that were previously linked to schizophrenia using ClinVar.

3.4.1 Structural variants on chromosome 22

Approximately 1% of patients with schizophrenia also have the 22q11.2 deletion syndrome (**22qDS**) [131]. It is assumed that the presence of large **Indels**, especially deletions, on chromosome 22 may contribute to the neurodevelopmental mechanisms involved in the pathogenesis of

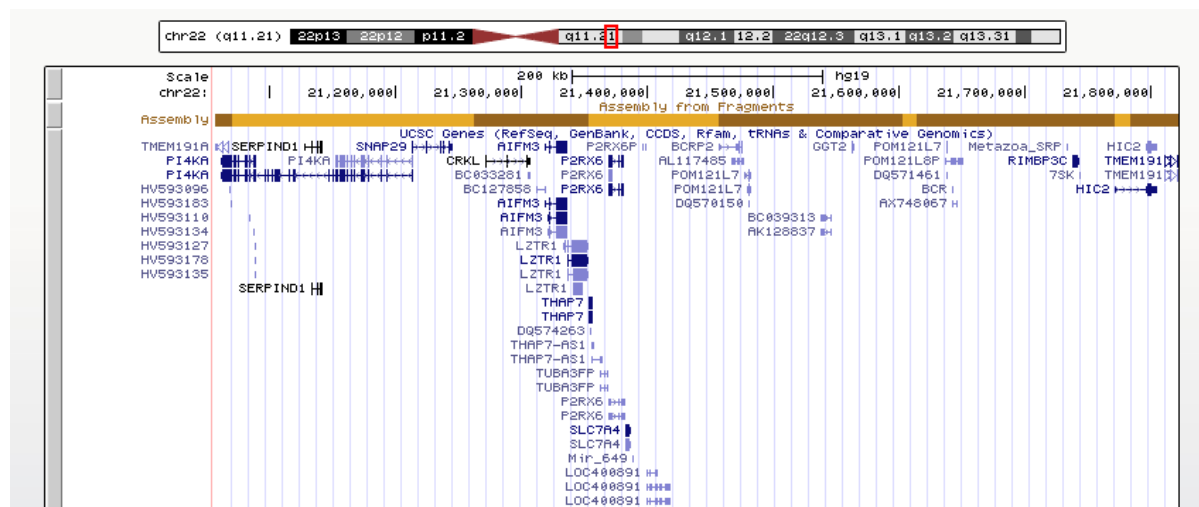


Figure 3.17: 765 kilo base pair deletion on chromosome 22 displayed in the UCSC Genome Browser [132]. As indicated by the red box, it affects the 22q11 region and is located close to the boundary between 22q11.21 and 22q11.22. In total, 14 genes are affected by the deletion which are annotated below the assembly (blue)

schizophrenia. Schizophrenia patients with **22qDS** typically have slightly different physical and auxiliary clinical features and can be seen as a special genetic subtype in the field of schizophrenia. Here, I used the Pindel, a tool for detecting breakpoints of large deletions, medium sized insertions, inversions, tandem duplications and other structural variants.

In total, I found 26 deletions with a length >100 base pairs. Among those are 21 deletions with a length >10.000 base pairs that are located within the 22q11 region (Table 3.4). The longest deletion with a length of 765 kilo base pairs affects a total of 14 genes (*AIFM3*, *CRKL*, *HIC2*, *LRRC74B*, *LZTR1*, *P2RX6*, *PI4KA*, *RIMBP3B*, *RIMBP3C*, *SERPIND1*, *SLC7A4*, *SNAP29*, *THAP7*, *TMEM191C*). As shown in Figure 3.17, the position is located close to the boundary between 22q11.21 and 22q11.22. Although the variant occurs heterozygous in the respective individual, a loss of hundreds of thousands base pairs will nevertheless cause drastic genetic effects as not only genes but also regulatory elements are affected.

3.4.2 ClinVar

ClinVar is a NCBI database, that aggregates information about genomic variation and its relationship to human health [105]. In order to see whether known, schizophrenia-associated mutations or genes were included in the dataset, I used this resource for selecting potential disease relevant variants. In total, this resulted in 545 putative candidate variants, located in 31 schizophrenia-associated genes. To discard low impact and potential unrelated mutations, I used the previously introduced frequency and deleteriousness cutoffs. The remaining 150 variants distributed among 25 genes whereas most mutations were found in *HSPG2* (Heparan Sulfate

3.4. ANALYSIS OF STRUCTURAL VARIANTS AND ASSOCIATION WITH KNOWN CANDIDATES

Proteoglycan 2) and *ABCA13* (ATP Binding Cassette Subfamily A Member 13). Altogether, several promising variants were identified which are shown in Table 3.5.

Gene	Mouse Phenotype	Parameter	Center	Ref.
AARS2	impaired pupillary reflex	Pupil Light Response	MARC	[115]
ALPK2	abnormal behavioral response to light	Side changes	JAX	[116]
ALPK2	abnormal behavioral response to light	Time mobile light side	JAX	[116]
ASPA	abnormal behavior	Center average speed	HMGU	[117]
ASPA	abnormal behavior	Center distance travelled	HMGU	[117]
ASPA	abnormal behavior	Number of center entries	HMGU	[117]
ASPA	abnormal behavior	Periphery average speed	HMGU	[117]
ASPA	abnormal gait	Gait	HMGU	[118]
ASPA	decreased food intake	Total food intake	HMGU	[119]
ASPA	decreased grip strength	Forelimb grip strength measurement mean	HMGU	[120]
ASPA	decreased prepulse inhibition	% Pre-pulse inhibition - Global	HMGU	[121]
ASPA	decreased prepulse inhibition	% Pre-pulse inhibition - PPI2	HMGU	[121]
ASPA	decreased prepulse inhibition	% Pre-pulse inhibition - PPI4	HMGU	[121]
ASPA	decreased startle reflex	Response amplitude - S	HMGU	[122]
ASPA	decreased vertical activity	Number of rears - total	HMGU	[123]
ASPA	hyperactivity	Periphery resting time	HMGU	[124]
ASPA	hyperactivity	Whole arena resting time	HMGU	[124]
ASPA	hypoactivity	Distance travelled - total	HMGU	[123]
ASPA	hypoactivity	Locomotor activity	HMGU	[123]
ASPA	hypoactivity	Whole arena average speed	HMGU	[123]
CDH23	abnormal gait	Gait (inc. ataxia)	WTSI	[118]
CDH23	absent pinna reflex	Startle response	WTSI	[122]
CDH23	stereotypic behavior	Head bobbing/circling	WTSI	[125]
DAGLA	abnormal behavior	Center average speed	MARC	[117]
DAGLA	abnormal behavior	Center permanence time	MARC	[117]
DAGLA	abnormal behavior	Percentage center time	MARC	[117]
DAGLA	abnormal behavior	Periphery permanence time	MARC	[117]
DAGLA	decreased brain size	Brain	MARC	[126]
DAGLA	decreased grip strength	Forelimb and hindlimb grip strength measurement mean	MARC	[120]
DLG5	abnormal brain morphology	Brain	UC Davis	[127]
DOCK6	abnormal sleep behavior	Sleep bout lengths standard deviation	JAX	[128]
DOCK6	hyperactivity	Periphery distance travelled	JAX	[124]
DPYD	abnormal behavior	Periphery average speed	TCP	[117]
ENTPD1	decreased grip strength	Forelimb grip strength measurement mean	HMGU	[120]
ENTPD1	decreased grip strength	Forelimb grip strength normalised against body weight	HMGU	[120]
ENTPD1	tremors	Tremor	MRC Harwell	[125]
GRHL3	abnormal locomotor activation	Body position	ICS	[125]
HHATL	hypoactivity	Locomotor activity	UC Davis	[123]
HSPB2	abnormal motor learning	Learning difference	JAX	[125]
LRBA	absent pinna reflex	Startle response	WTSI	[122]
LRRCC1	abnormal brain morphology	Brain	UC Davis	[127]
MACROD2	abnormal behavioral response to light	Time mobile dark side	JAX	[116]
MACROD2	abnormal behavioral response to light	Time mobile light side	JAX	[116]
MACROD2	abnormal sleep behavior	Peak wake with respect to dark onset median	JAX	[128]
MACROD2	convulsive seizures	mA threshold inducing clonic seizure	JAX	[129]
MYO9A	decreased food intake	Total food intake	WTSI	[119]
NAV2	abnormal behavior	Center distance travelled	TCP	[117]
NAV2	abnormal behavior	Latency to center entry	TCP	[117]
NAV2	abnormal behavioral response to light	Time mobile light side	TCP	[116]
NAV2	hyperactivity	Periphery resting time	TCP	[124]
NAV2	hyperactivity	Whole arena resting time	TCP	[124]
NAV2	hypoactivity	Locomotor activity	TCP	[123]
NEK1	decreased grip strength	Forelimb and hindlimb grip strength measurement mean	BCM	[120]
NEK1	decreased grip strength	Forelimb grip strength measurement mean	BCM	[120]
NEK1	decreased grip strength	Forelimb grip strength normalised against body weight	BCM	[120]
NEURL4	decreased prepulse inhibition	% Pre-pulse inhibition - PPI1	UC Davis	[121]
NEURL4	decreased prepulse inhibition	% Pre-pulse inhibition - PPI2	UC Davis	[121]
NOTCH3	decreased prepulse inhibition	% Pre-pulse inhibition - PPI3	UC Davis	[121]
NRAP	increased anxiety-related response	Fecal boli	JAX	[130]
OLFML2B	abnormal behavior	Center distance travelled	JAX	[117]
OLFML2B	abnormal behavior	Center permanence time	JAX	[117]
OLFML2B	abnormal behavior	Percentage center time	JAX	[117]
OLFML2B	abnormal behavior	Periphery permanence time	JAX	[117]
OLFML2B	abnormal sleep behavior	Sleep daily percent	JAX	[128]
PARD3B	abnormal behavior	Center resting time	HMGU	[117]
PARD3B	hyperactivity	Whole arena resting time	HMGU	[124]
PKHD1L1	decreased startle reflex	Response amplitude - S	MRC Harwell	[122]
PPARG	abnormal gait	Gait	MRC Harwell	[118]
PRODH	abnormal vocalization	Vocalization	UC Davis	[123]
SORL1	abnormal gait	Gait	MRC Harwell	[118]
SPNS3	abnormal behavior	Center average speed	TCP	[117]
TRPM1	impaired pupillary reflex	Pupil Light Response	MRC Harwell	[115]
USP31	abnormal sensory capabilities	% Pre-pulse inhibition - PPI1	MARC	[121]
USP31	decreased prepulse inhibition	% Pre-pulse inhibition - PPI2	MARC	[121]
USP31	hypoactivity	Periphery resting time	MARC	[123]
USP31	hypoactivity	Whole arena resting time	MARC	[123]
WDR37	decreased grip strength	Forelimb grip strength measurement mean	WTSI	[120]

Table 3.3: Candidate genes with significant schizophrenia-related mouse phenotypes ($P < 0.0001$). Along with the detected phenotype, the measurement parameter, the phenotyping center and evidence for links to schizophrenia is provided. Data obtained from the [IMPC](#) [107].

3.4. ANALYSIS OF STRUCTURAL VARIANTS AND ASSOCIATION WITH KNOWN CANDIDATES

Chr.	Genomic Pos.	Type	Affected Gene(s)	Effect	Location	Length
chr22	21372403	exonic	<i>P2RX6, SLC7A4</i>	frameshift deletion	22q11.21	24.166
chr22	22724137	intergenic	<i>BMS1P20, ZNF280B</i>	-	22q11.22	25.373
chr22	23101160	upstream	<i>MIR650</i>	-	22q11.22	64.085
chr22	22707572	intergenic	<i>BMS1P20, ZNF280B</i>	-	22q11.22	74.590
chr22	23101701	exonic	<i>IGLL5</i>	frameshift deletion	22q11.22	140.098
chr22	25625383	exonic	<i>CRYBB2, LRP5L</i>	frameshift deletion	22q11.23	227.822
chr22	21057064	exonic	<i>AIFM3, CRKL, HIC2, LRRC74B, LZTR1, P2RX6, PI4KA, RIMBP3B, RIMBP3C, SERPIND1, SLC7A4, SNAP29, THAP7, TMEM191C</i>	frameshift deletion	22q11.21	765.322

Table 3.4: Long deletions detected by Pindel that affect the 22q11 region. For each deletion, the affected genes are specified including the exact region and length of the variation.

Chr.	Genomic Pos.	Mut.	CADD	rsID	VariantAssoc	het/hom	Gene	Type	Exon	AF
7	48349706	C/T	39	rs376273529	-	1/0	ABCA13	stop gained	24/62	0.001
7	48450185	C/G	25.2	rs77147473	-	7/1	ABCA13	NS	40/62	0.0229
17	27286356	G/A	42	-	-	1/0	SEZ6	stop gained	9/17	8.668×10^{-5}
22	18905859	G/A	23.2	rs2870984	Schizophrenia	2/0	PRODH	NS	11/14	0.0201
22	18905899	G/A	28.9	rs3970559	Schizophrenia	9/0	PRODH	NS	11/14	0.0401
22	20230079	C/T	27.2	-	-	1/0	RTN4R	NS	2/2	-1
22	51117236	C/T	25.6	rs377609278	-	1/0	SHANK3	NS	5/23	0.0002

Table 3.5: Excerpt of rare deleterious variants that were found to be associated with schizophrenia, either directly or by gene association. **het/hom** column specifies the number of cases with a heterozygous or homozygous genotype; **Exon** indicates the affected exon; **AF** is the maximal allele frequency across ExAC,1000G and ESP

DISCUSSION

4.1 Comprehensive whole-exome sequencing analysis of sporadic schizophrenia

I have analysed a whole-exome data set of 37 schizophrenia trios and 210 further patients. By applying a novel combination of multiple methods that cover various aspects of the mutational spectrum, I obtained a comprehensive set of putative candidate genes involved in the causation of schizophrenia. These include genetic variants caused by *de novo* events but also inherited mutations that might contribute to predisposition to schizophrenia.

In the following, I will discuss these results from a biological point view and will put them into the broader context of schizophrenia and mental health.

4.1.1 Rate of *de novo* variants supports important role in schizophrenia

De novo variants are well known to strongly contribute to the genetics of schizophrenia [50, 51, 56]. As shown in Table 3.1, a total of 46 mutations remained in the high confidence set of rare deleterious variants. Overall, this implicates an increased rate of *de novo* mutations (1.8×10^{-8}) within this dataset which is in accordance with previous studies [49, 51]. Besides being more frequent, several variants were found to be located in schizophrenia-related regions.

Notably, three previously schizophrenia-associated loci SLC45A1, MYO1A and TMEM219 [112] were found. Additionally, 13 of 46 affected genes were found to have a direct protein-level interaction to at least one of the genes located within the 108 schizophrenia loci detected by the Schizophrenia Working Group of the Psychiatric Genomics Consortium [112].

Further notable genes include:

SYNE1 (Spectrin Repeat Containing Nuclear Envelope Protein 1) a protein coding gene that is expressed in skeletal and smooth muscle, and peripheral blood lymphocytes, and localizes to the nuclear membrane. A recent case-control study concluded that polymorphisms in SYNE1 may confer a greater risk of developing schizophrenia, especially for individuals with schizophrenia family history [133]. Here, I found a *de novo* variant shifting the nucleotide reading frame of the gene. Based on the available data and literature, SYNE1 seems to be a promising target for further investigations as it may be one of the genetic factors contributing to schizophrenia risk.

HTR4 (5-Hydroxytryptamine Receptor 4) a member of the family of serotonin receptors which stimulate cAMP production in response to serotonin. This is of special interest, as the serotonin hypothesis of schizophrenia has a long history [134]. Also serotonin receptors have been postulated to play a critical role in schizophrenia drug development [135].

NCG Next to the directly schizophrenia-linked genes, several Neuronal Cytoskeleton-associated Genes (**NCG**) were detected during this analysis. This especially includes STRN (Striatin, involved in dendritic Ca(2+) signaling), DST (Dystonin, retrograde axonal transport) and MYO1A (myosin IA, transport along actin filaments). All of these will be discussed in Section 5.2.

Taken together, these findings support the prominent role of *de novo* variants in schizophrenia and highlight the importance of trio/family-based sequencing.

4.1.2 Rare deleterious compound heterozygous variants affect a broad range of genes

Disruptive compound heterozygous variants have been demonstrated to underlie many disorders including some complex psychiatric disorders such as autism [136]. Although so far no major role of compound heterozygous variants in schizophrenia was found, they still are assumed to play a certain role in disease risk [136]. Still, in many studies compound effects were neglected so far. In order to focus on the most disease-relevant variants, only rare compound heterozygous variants were considered for the final candidate list. Overall, this included 354 genes that were affected by at least one rare compound heterozygous variant. Among others, this includes several promising targets:

NRG1 (Neuregulin-1) mediates cell-cell signaling and plays crucial role in the development of multiple organ systems. Throughout the recent years, several studies indicated an association with schizophrenia. Although, so far, no genome-wide significant NRG1 variant was found, the combination of findings resulting from linkage analyses, family and case-control

studies indicate the relevance NRG1 in schizophrenia [137]. While NRG1 is arguably one of the most extensively characterized schizophrenia candidate genes, the exact functions are not well understood. Still, there is certain evidence for NRG1 controlling neurotransmission via the ErbB4 receptor with downstream effects on cognitive processes involving frontal cortex, hippocampus and striatum [137].

CMYA5 (Cardiomyopathy-associated gene 5) encodes myospryn, a large tripartite motif (TRIM)-related protein which is primarily known to be associated with diseases affecting striated muscle. Still, several studies proposed that CMYA5 should be considered as a schizophrenia risk gene [138, 139]. The interaction with DTNBP1, another proposed schizophrenia candidate with physiological significance in the central nervous system, further supports this case [140].

NEK1 (NIMA-related kinase 1) functions in multiple ways in neurons including maintaining the neuron's cytoskeleton and axonal development and plays a role in centrosome integrity, affecting both ciliogenesis and centrosome stability. Although NEK1 is mostly known as a prominent candidate in Myotrophic Lateral Sclerosis (ALS) [141], it was also identified as a potential schizophrenia-associated locus [112]. Interestingly, a recent study by McLaughlin et al. concluded that schizophrenia and ALS share neurobiological mechanisms to a higher-than-expected rate [142]. Both findings support a potential role of NEK1 in schizophrenia as well as the neuronal cytoskeleton which will be discussed in Section 5.2 along with other NCGs.

In addition to those presented candidates, further 16 genes (CCDC39, CHAT4, CR1L, DPYD, EPB41L2, GSN, ITPR3, JMJD1C, MAGI3, MYO18B, PDCD11, PLXNA2, PRODH, RPGRIP1L, SLC9B1, SRMS) can be directly linked to schizophrenia throughout literature and database research. Although it is unlikely that all 354 genes with rare deleterious compound heterozygous variants actually contribute to the disease as they might also occur by chance, the underlying candidate gene selection based on rare variants allows for an enrichment of such, especially when assuming that mostly rare alleles are responsible for disease risk.

4.1.3 Rare potential LoF variants unveil genes affected by genetic knockout

Genetic variants that disrupt protein-coding genes, are known as LoF variants. Unlike most other variant types, they have been regarded as rare and highly deleterious. However, several studies showed that even healthy individuals carry up to 100 protein-truncating variants, although the exact numbers vary [143, 144]. Most of those variants result in a heterozygous LoF, leaving one copy of the allele active. There are also several common LoF alleles which are present in a large fraction of the population. Typically, a loss of those genes will only have a weak or neutral effect on health and overall fitness [143]. This leaves a smaller fraction of Rare Homozygous

Loss-of-Function (**rhLoF**) genotypes that tend to lead to more severe phenotypic effects due to the genetic knockouts.

In my work, I found about 140 potential **LoF** variants per individual, which is a slight increase to published numbers by MacArthur et al. [144] but is in accordance to a recent investigation by Narasimhan et al. [145]. In this study, the authors detected 140.3 predicted **LoF** genotypes per sample in a cohort of 3,222 healthy British Pakistani-heritage adults [145].

While most of the variants found in the m4-schizophrenia data set are heterozygous state, a smaller fraction of about 30 variants affects both alleles. Not all of these consequently lead to full knock-out of the gene product since, depending on the actual mutation site, not all gene transcripts might be disrupted. Overall, only about 10 genes were found as completely inactive across all transcripts. For rare variants, this number drops even further to about 5 genes per individual. This low number comes as no surprise as this is one of the most severe mutational changes that can be introduced into the genomic DNA. No difference between healthy parents and affected cases was found, indicating that there is no general increase of **LoF** variants in schizophrenia within this cohort.

Based on those results, I extracted all **rhLoF** genotypes as well as compound heterozygous combinations of **LoFs** that occur exclusively in affected subjects and are therefore not present in the respective parents. Overall, this leads to 10 distinct candidate genes.

Unfortunately, a carefully performed literature investigation resulted in no clear associations to schizophrenia or neuronal cytoskeletal candidate genes which most likely can be explained by the very small number of candidates.

4.1.4 Genetic burden analysis reveals new targets with increased genetic load

The analysis of the genetic burden is an unsupervised approach to discover individual or groups of genetic loci that harbour deleterious variants. In this specific setup, each offspring was directly compared to the respective parents to determine the relative shift of the genetic load. Such changes can be introduced by two possibilities, genetic recombination as well as *de novo* events. Both cases hold the possibility to create a deleterious combination of variants that lead to severe downstream effects.

Overall, I detected 35 candidate genes that showed an increased genetic burden in more than 1/3 of the investigated trio offsprings.

Among others, this includes some schizophrenia-associated loci as well as the previously mentioned **NCGs**.

ITIH4 (Inter-Alpha-Trypsin Inhibitor Heavy Chain Family Member 4) is a known schizophrenia candidate that has been identified by recent GWAS studies in European and Han Chinese populations [146]. Further investigations supported these findings of associations of genetic

variants in ITIH4 with schizophrenia concluding a potential role of ITIH4 in the risk for schizophrenia throughout genetic gene expression control [147].

MYO18B (myosin XVIIIIB) is another well known schizophrenia-associated candidate that has been found during a large GWAS performed the International Schizophrenia Consortium [148].

NCG As before, also this analysis yielded several **NCGs** (KIF13A,TUBB8, DNAH7, DNAH8) that will be discussed later on (Section 5.2).

4.1.5 Edgetics provides insights to the effect of genetic variants on PPI networks

For schizophrenia and most typical complex diseases the model of one-gene/one-disease is clearly not applicable. Instead, the prevalence of complex genotype-to-phenotype associations that often include multiple genes or gene groups is widely accepted [149]. Since those genes and their products typically don't act in an isolated environment but are rather in a continuous interaction to other biochemical components, the investigation of protein interactions is of particular interest. Those interactions, the interactome, can be represented as a network model in which genes are typically shown as nodes, whereas each edge depicts a physical interaction between two network components. Thus, the study of edgetics investigates the loss or gain of interactions in order to draw conclusion for genotype-to-phenotype relationships [96].

As shown in Section 3.2.5, I detected 31 interrupted **PPI** interface sites that were disrupted by rare genetic variants. Interestingly, about 2/3 of affected **PPIs** include self-interacting proteins. Considering the fact that only about 41% of the underlying database interactions represent such homomeric interactions, this clearly indicates an over-representation of this interaction type (Fisher's exact test: 8.93×10^{-3})

In contrast to heteromeric interactions, in which only one interaction site is mutated, a homomeric interaction can be affected twice if the same interaction site is used by both partners, resulting in an even larger effect. A reason for this unexpected observation might be that due to the potentially larger impact, mostly rare variants are found in such 'double-mutant' sites.

In total, the 31 variants affect 37 unique genes (28 mutated proteins and 9 additional interaction partners). Compared to previous analysis steps, only a limited amount of genes with a biological connection the schizophrenia were found. Foremost this includes CRYAB (Crystallin Alpha B), a member of the small heat shock protein (HSP20) family. Elevated expression of alpha-B crystallin has been found in many neurological diseases and was shown by Lin et al. for schizophrenia and autism candidate genes [150]. In addition, CRYAB is also involved in microtubule binding which will be discussed in Section 5.2.

4.1.6 Integrative analysis and structural variation

The integration of information coming from multiple sources offers the possibility to create a broader overview across the analysed data. In the course of this analysis, I used multiple tools and approaches for generating new insights on the present data. In particular, this includes the integration of PPIs, phenotypic information gained by mouse knock-out models, variants and genes with annotated associations to schizophrenia (ClinVar [105]) as well as the analysis of structural genomic variants using Pindel [104].

4.1.6.1 Protein hubs and network interconnectivity

As described in Section 3.3.1 and shown Figure 3.13, the PPI network consisting out of 468 candidate genes and their direct interaction partners, provides several strongly connected proteins. These molecules tend to interact with a large number of surrounding proteins, often sharing common pathways or function. Obviously, those network hubs are often crucial in their function as an interconnecting element of the whole biological system and therefore are of special interest when affected by deleterious mutations. Among the network hub proteins, nine were included in the previously created candidate list. Prominently, the network also included UBC (Ubiquitin C) which accounts for almost 20% of all network connections. This comes as no surprise as the diversity of UBC contributes to the regulation of many cellular events across the human biological system [151]. When reducing the network to previously identified candidates, the network showed no significant enrichment of connections. This finding indicates that the candidates are spread across multiple processes and not directly enriched for specific pathways. Nevertheless, the inclusion of first degree interaction partners showed a second level relatedness and interconnectivity of the network proteins.

Among the nine candidate protein hubs especially STRN, CUL7, ANAPC1 and NPAS2 seem promising targets for further investigation. Next to their centrality in the gene interaction network, those genes have been found to take part in either schizophrenia or neurodevelopmental processes. Striatin (STRN) is assumed to play a role in dendritic Ca(2+) signaling and together with PP2A modulates microtubule dynamics [152]. Also Cullin 7 (CUL7) alters microtubule dynamics [153] and plays a role in neural dendrite patterning and growth [154]. The third gene, Anaphase Promoting Complex Subunit 1 (ANAPC1, APC1) has been shown to be essential for maintenance of local brain organizers and midbrain survival as a knockdown of Apc1 resulted in reduced proliferation and apoptosis in the dorsal midbrain [155]. The last identified target is the Neuronal PAS Domain Protein 2 (NPAS2) which was previously associated with schizophrenia and schizoaffective disorder [156].

4.1.6.2 Knock-out mouse phenotyping reveals links to schizophrenia

Phenotyping using knock-out mice is a well-established way for uncovering so far unknown gene functions. To investigate whether known relevant phenotypic links for the detected candidate genes exist, I relied on data generated by the [IMPC](#). As shown in Section [3.3.3](#) and Table [3.3](#), several candidates were already linked to schizophrenia-related phenotypes.

Out of 36 genes with available phenotypic information, the associated phenotypes for 32 genes were directly linked to schizophrenia by literature evidence (Table [3.3](#)). This included several promising findings:

DAGLA (Neural Stem Cell-Derived Dendrite Regulator) encodes a diacylglycerol lipase, highly expressed in the brain, that is required for axonal growth during development and for retrograde synaptic signaling at mature synapses [[157](#)]. Based on [IMPC](#) data, a knockout of DAGLA leads to abnormal behavior, decreased brain size and decreased grip strength. All three phenotypes have been shown to be linked to schizophrenia [[117](#), [120](#), [126](#)]. In this study, the gene was found to have an increased genetic burden in cases compared to healthy parents.

DLG5 (Discs Large MAGUK Scaffold Protein 5) encodes a member of the family of discs large (DLG) homologs. It was found to regulate dendritic spine formation and synaptogenesis in mice [[158](#)]. According to [IMPC](#) data, a heterozygous knockout of Dlg5 leads decreased brain size whereas a homozygous knockout was found to be lethal. In this study, one case carried a compound heterozygous combination of rare deleterious variants that could strongly affect the genes function.

NEK1 (NIMA Related Kinase 1) encodes a serine/threonine kinase involved in cell cycle regulation. It interacts with FEZ1 [[159](#)] that is assumed to mediate microtubule-based cargo transport [[160](#)]. [IMPC](#) data showed decreased grip strength in homozygous knockout mice. Grip strength was found to be significantly associated with cognitive functioning in individuals with schizophrenia, particularly for working memory and processing speed [[120](#)]. As DLG5, also NEK1 was affected by a compound heterozygous combination of rare deleterious variants in one case.

NCG Also for several [NCGs](#) phenotypic links were found (CDH23, MYO9A, NERUL4, NOTCH3, NRAP) that will be discussed later on (Section [5.2](#))

4.1.6.3 ClinVar and known schizophrenia-associated genes

Besides rare mutations, frequent genomic variants are found across many studies investigating various specific phenotypes. This valuable information is gathered within publicly available variant databases. In my study, I relied on ClinVar, a database for information about genomic variation and its relationship to human health. My analysis focused on known associations with

schizophrenia and schizophrenia-related terms. As mentioned in Section 3.4.2, several promising candidates were detected which I will discuss in the following.

In total I detected 150 rare damaging variants that distributed among 25 genes. About one third of these variants are located within two specific genes, *ABCA13* and *HSPG2*. I identified 23 rare variants across *ABCA13*, a susceptibility factor for both schizophrenia and bipolar disorder as shown by Knight et al. [161] as well as 25 in *HSPG2*, which has been associated with childhood-onset schizophrenia. Furthermore, I detected two non-synonymous SNP (rs2870984, rs3970559) in *PRODH* (Proline Dehydrogenase 1), located in the 22q11.21 region, that have been associated with schizophrenia susceptibility. Also, the chromosomal 22q11.2 region itself is a target of excessive research that is known as a strong genetic risk factor for schizophrenia which will be discussed later on.

When considering the accumulation of variants in *ABCA13* and *PRODH*, one has to keep in mind that both genes exceed the average human gene length of about 8.446 bp (s.d. = 7124) [162]. With a total length of 476.050 bp, *ABCA13* is currently the largest known gene of the ABC family [163], while also *PRODH* comes up with 23.859 base pairs. Therefore, especially the amount of variants found in *ABCA13*, is more likely due to the total gene length than due to the phenotypic background. Nevertheless, given previous findings[161], a certain functional impact of these variants seems to be plausible.

Another finding of relevance is the detected LoF variant in SEZ6. Seizure Related 6 Homolog (SEZ6) is annotated as a childhood-onset schizophrenia risk gene that is necessary for balance between dendrite elongation and branching during the elaboration of a complex dendritic arbor. Furthermore it is assumed to be involved in the development of appropriate excitatory synaptic connectivity [164]. Function, variant type and previous study results clearly suggest an increased impact of this variant on neuronal functions, e.g. the neuronal cytoskeleton (See Section 5.2).

4.1.6.4 The 22q11.2 deletion and its association to schizophrenia

Until today, the strongest known molecular genetic risk factor for schizophrenia is the 22q11.2 deletion syndrome. Schizophrenia occurs in about 25% of cases that were diagnosed with this deletion located close to the center of chromosome 22 [165].

Due to this strong association between both phenotypes, I analysed chromosome 22 for the presence of long deletions. As shown in the results section, I did not detect any homozygous deletion that spanning the relevant positions. This comes as no surprise due to the low prevalence of the 22q11.2 deletion in the general population (1/4000-1/6000) [166]. In addition, no 22q11.2 deletion phenotype was noted during sample recruitment, supporting this observation. Still, the strong association between the 22q11.2 and schizophrenia implicates the assumption that several schizophrenia related genes are located within this genomic region. Therefore, also deletions that only partially affect this specific locus might be of disease-relevance.

As shown in Table 3.4, especially the heterozygous deletion starting at position 21,057,064 (Fig-

ure 3.17) which is located close to the boundary between 22q11.21 and 22q11.22 seems to be a promising candidate. It affects a total of 765 kilo bases including 14 genes. Among others these includes genes like CRKL, LZTR1 and PIK4CA that all have been found to play important neurological roles [167–169].

Especially CRKL (CRK Like Proto-Oncogene, Adaptor Protein), a gene that is known to be haploinsufficient, is associated with dysfunction of the neural crest cells when being affected by mutation i.e deletion of one copy [167].

LZTR1, another gene included within the 765kb deletion, encodes a protein exclusively localized to the Golgi network helping to stabilise the Golgi complex [170]. Previous research investigating LZTR1 and the Golgi complex established a strong association between Golgi apparatus and neuronal phenotypes caused by the 22qDS [171]. PIK4CA (phosphatidylinositol-4-kinase-catalytic- α) is a gene that was repeatedly reported harbouring schizophrenia-associated SNPs concluding that it may be involved as one of various genetic factors that contribute to schizophrenia [169]. Altogether, this analysis showed that within the 22q11.2 region multiple genes with neurological functions can be found in the analysed cohort. On the one hand, this supports the causal relation to the 22qDS and on the other hand, the strong association between 22qDS and schizophrenia.

4.1.7 Variant distribution and frequency of deleterious events

In the previous sections, I discussed multiple distinct variant types and their effects on genes and pathways. In the following, I will provide an overview about the distribution of these variants across the analysed samples and discuss the overall frequency of the mutation events detected compared to the neutral expectation and in the context of recent literature.

As shown in Figure 4.1, the distribution of the mutation events largely varies depending on the respective type. A new genetic knockout of a gene caused by LoF variants typically occurs in about one third of all trio samples. Multiple knockouts were only seen rarely. Given that the overall frequency of LoF variants is in accordance with literature [145], it can be assumed that the amount of *de novo* knockouts, that are not present in the respective parents, does not deviate from the expectation.

The situation changes when considering *de novo* variants. Here, as discussed earlier (Section 4.1.1), the frequency of *de novo* events is significantly increased compared to the neutral expectation. Accordingly, rare deleterious *de novo* mutations can be found in about 75% of all trio samples. This increase of *de novo* mutations has been described earlier by multiple studies [49, 51].

In contrast to the LoF and *de novo* variants, a statement about compound heterozygous variants cannot be made easily. Here, various parameters directly influence the final set of variants. First and foremost, the calling of compound heterozygous variants relies on phased genotypes. Since phasing of non-trio samples is often a complex process and not always possible, a clear reference value for the frequency of compound heterozygous variants can not be obtained from the data or literature. Also, since the variants are not easily detectable, it can be assumed that the false

negative rate is increased. Instead, a control cohort would be required to compare the values on a common basis. Nevertheless, it can be noted that rare deleterious compound heterozygous variants are found in almost all samples, with most of them carrying more than one.

Variants affecting PPI sites were found in about one third of all samples. Based on the total number of variants found in the schizophrenia cohort, the number of known PPI interfering variants and the number of bases in the sequencing capture kit, this is a significant increase compared to the neutral expectation ($p=2.281 \times 10^{-13}$). This is surprising as one would assume that protein interaction sites should be highly conserved and under strong purifying selection. Still, the observation should be validated using an independent sample set using processed under similar conditions.

In contrast to the Edgetics approach, no reference database is available when considering genes with increased genetic burden. Since this approach is based on a custom pipeline that was developed during this work, no references are available. Therefore no conclusion about the frequency of these events can be drawn without an appropriate control cohort. Concerning absolute numbers, about 50% of all trio samples carried between 10 and 15 genes that showed an increased genetic load compared to their healthy parents.

The last category to be discussed, are the known variants annotated in ClinVar. In my study, slightly more than 50% of all affected samples carry at least one variant associated with schizophrenia or related terms. This, however, is not surprising as these annotated variants are not required to be rare. Consequently, common variants are included that exist in healthy individuals, too.

Summarising on this point, it is clear that in order to obtain reliable statements about deviating frequencies, a well-selected control cohort is necessary. Only with comparable underlying data processing, accurate conclusions can be made. Still, the increased rate of *de novo* variants seems reasonable in the context of previous published research. A full sample-specific overview providing exact counts for all variation types is shown in Table 8.1.

4.1.8 Enrichment Analysis

In order to gather a comprehensive set of candidate genes that are putatively linked to causative pathways of schizophrenia, I performed various distinct analyses covering multiple aspects of the mutational spectrum. As a result, the list of candidates quickly grows to an extent that is no longer feasible for manual inspection. Therefore, dedicated gene enrichment methods can be used to make sense out of genes lists in an effort to gain a more general data insight. Accordingly, I used two different approaches (WebGestalt and DNENRICH) to detect overrepresented biological pathways and groups.

4.1. COMPREHENSIVE WHOLE-EXOME SEQUENCING ANALYSIS OF SPORADIC SCHIZOPHRENIA

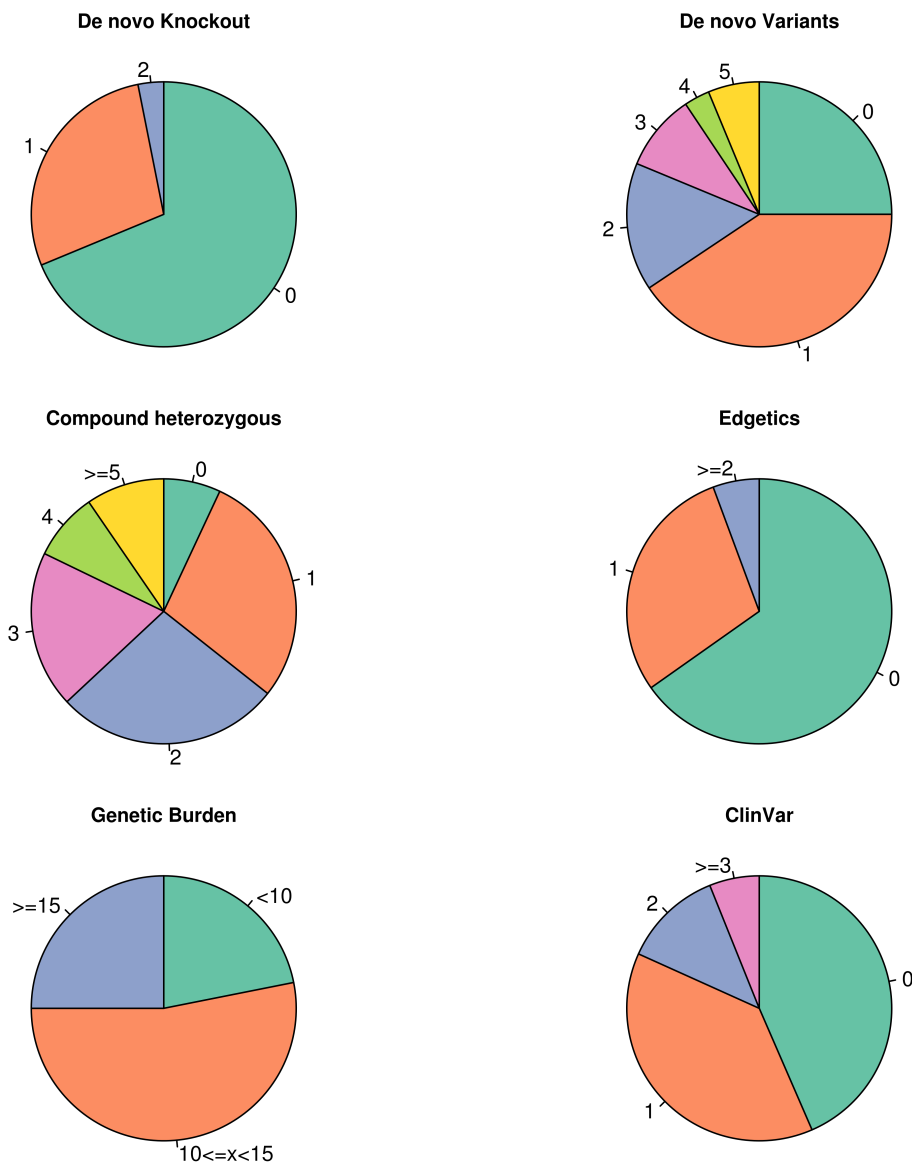


Figure 4.1: Distribution and frequency of six variant types among schizophrenia samples. For each investigated type, the amount of cases is indicated that is affected by the respective number of events. For *de novo* knockout (LoF), *de novo* Variants, Edgetics and ClinVar this event corresponds to the amount of actual SNPs or Indels that were found. For compound heterozygous variants the number indicates the amount of heterozygous variant pairs that were detected as compound heterozygous. In case of the genetic burden analysis, the values indicate the number of genes with increased genetic burden. Subplots for *de novo* knockout, *de novo* variants and genetic burden are based on trio cases only. Compound heterozygous, Edgetics and ClinVar include the complete cohort.

4.1.8.1 Significant GO categories and their relevance for schizophrenia

In total, 16 GO terms across all three categories (Biological process, Molecular function, Cellular component) were found with a significant enrichment of candidate genes. In the following, I will briefly discuss the categories in order of decreasing enrichment significance (WebGestalt). All GO term definitions were retrieved from AmiGO2 [114].

Motor activity (GO:0003774)

Def.: *"Catalysis of the generation of force resulting either in movement along a microfilament or microtubule, or in torque resulting in membrane scission, coupled to the hydrolysis of a nucleoside triphosphate."*

Pvalue WebGesalt: 5.97×10^{-13} , **Pvalue** DNENRICH: 1.13×10^{-4}

The motor activity category includes a large variety of genes since inter- and intra-cellular cargo transport requires the involvement of molecular motors. The three main classes of molecular motors, the kinesin, dynein, and myosin superfamily all play fundamental roles in neuronal function, plasticity, morphogenesis and transport of synaptic vesicle precursors, neurotransmitter and neurotrophic factor receptors within axons, dendrites, and synapses making them a critical component involved in neuronal disease pathogenesis [172]. Here I found a total of 25 unique motor proteins distributed among all three molecular motor superfamilies, making this the top enrichment hit with a strong connection towards neuronal malfunction.

Actin binding (GO:0003779)

Def.: *"Interacting selectively and non-covalently with monomeric or multimeric forms of actin, including actin filaments."*

Pvalue WebGesalt: 3.79×10^{-9} , **Pvalue** DNENRICH: 1.13×10^{-4}

Microfilaments, also called actin filaments, are filaments in the cytoplasm of eukaryotic cells that form part of the cytoskeleton. They are found in essentially all eukaryotic cells. As a result, actin binding proteins are not necessarily involved in any neurological or schizophrenia-relevant pathways. Still, a certain part of actin binding proteins plays a major role in dendritic spine formation, morphology, and function. These dendritic spines are actin-rich protrusions that comprise the postsynaptic sites of synapses and receive the majority of excitatory synaptic inputs in the central nervous system [173]. The actin binding proteins that were found to be disrupted within this study include various proteins with important neuronal functions like spectrin, dystonin and myosin as well as TRIOBP that has been implicated to be an element of the neuropathology of a multiple of chronic mental illnesses like schizophrenia [174].

Contractile fiber (GO:0043292)

Def.: "Fibers, composed of actin, myosin, and associated proteins, found in cells of smooth or striated muscle."

Pvalue WebGesalt: 5.14×10^{-6} , **Pvalue** DNENRICH: 2.04×10^{-4}

Due to the wide presence of cytoskeletal elements in eukaryotic cells and proteins that come with basic functions in various cell types, it comes as no surprise that also potentially unrelated gene groups are detected during enrichment analysis. In this case, the enrichment is more likely to be explained by the overlap between this and the previously discussed motor activity and actin binding classes. Genes unique to this biological gene set are less likely to be involved in neuropathology of schizophrenia although it can not be ruled out in general.

Cilium (GO:0005929)

Def.: "A specialized eukaryotic organelle that consists of a filiform extrusion of the cell surface and of some cytoplasmic parts. Each cilium is largely bounded by an extrusion of the cytoplasmic (plasma) membrane, and contains a regular longitudinal array of microtubules, anchored to a basal body."

Pvalue WebGesalt: 2.96×10^{-4} , **Pvalue** DNENRICH: 1.13×10^{-4}

Although neural cilia have been known to play a role in embryonic central nervous system patterning, they have been overlooked for a long time. However due to recent scientific progress, the cilium, especially the primary cilium, has gained attention within the field of neuroscience. Results suggest that neurons may also sense and respond to their environment via the primary cilium [175]. Furthermore, primary cilia formation was found to be diminished in schizophrenia and bipolar disorder [176]. Here we found 35 genes associated to the cilium. Among others this includes genes like RPGRIP1L that have been directly associated to schizophrenia [177].

ATPase activity (GO:0016887)

Def.: "Catalysis of the reaction: $ATP + H_2O \longrightarrow ADP + \text{phosphate} + 2H^+$. May or may not be coupled to another reaction."

Pvalue WebGesalt: 1.93×10^{-3} , **Pvalue** DNENRICH: 3.97×10^{-2}

Similarly to the microtubule associated complex, also the enrichment of ATPase activity can partially be explained by the frequently occurring motor proteins (14 out of 26) since they usually require force-generating ATPase activity for performing cellular transport processes [178]. Another group of genes classified into this set are ATP binding cassette (ABC) subfamily members. Although connections between ABC proteins and schizophrenia are known [179], the GO set is very unspecific providing little evidence for specific functional pathways.

Actin cytoskeleton (GO:0015629)

Def.: *"The part of the cytoskeleton (the internal framework of a cell) composed of actin and associated proteins. Includes actin cytoskeleton-associated complexes."*

Pvalue_{WebGesalt}: 1.44×10^{-3} , **Pvalue**_{DNENRICH}: 6.11×10^{-4}

In contrast to the previous group, the actin cytoskeleton defines a clear genetic landscape that is known to play fundamental roles in establishment and maintenance of both dendritic arborizations and spines. Alterations in the regulation of actin at dendritic spines have been shown to produce deficits in both synaptic plasticity and the formation and consolidation of long-term memories [180]. Although actin-specific motor and actin-binding proteins are present within this set of genes, various additional genes are included that further suggest disruptions in the overall cytoskeletal system. Among others this includes CGNL1 and CRYAB which have both been brought into context of schizophrenia neuropathology [150, 181].

Extracellular matrix structural constituent (GO:0005201)

Def.: *"The action of a molecule that contributes to the structural integrity of the extracellular matrix."*

Pvalue_{WebGesalt}: 1.93×10^{-3} , **Pvalue**_{DNENRICH}: 2.07×10^{-2}

On a first view, the term extracellular matrix seems to be not directly connected towards schizophrenia or mental health in general. However, as shown by Berretta et al. brain extracellular matrix abnormalities may contribute to several aspects of the pathophysiology of schizophrenia, including disrupted connectivity and neuronal migration, synaptic anomalies and altered GABAergic, glutamatergic and dopaminergic neurotransmission [182]. Interestingly, the most prominent gene family within this set are various forms of collagen. The loss of a prominent member of this family, collagen XIX, results in decreased inhibitory synapse number and the acquisition of schizophrenia-related behaviors [183]. Although disruptions in collagen XIX were not found, the enrichment disruptions of multiple types (IV, VI, IX and XII) of collagen are noteworthy.

Microtubule (GO:0005874)

Def.: *"Any of the long, generally straight, hollow tubes of internal diameter 12-15 nm and external diameter 24 nm found in a wide variety of eukaryotic cells; each consists (usually) of 13 protofilaments of polymeric tubulin, staggered in such a manner that the tubulin monomers are arranged in a helical pattern on the microtubular surface, and with the alpha / beta axes of the tubulin subunits parallel to the long axis of the tubule; exist in equilibrium with pool of tubulin monomers and can be rapidly assembled or disassembled in response to physiological stimuli; concerned with force generation, e.g. in the spindle."*

Pvalue_{WebGesalt}: 5.03×10^{-3} , **Pvalue**_{DNENRICH}: 5.28×10^{-4}

4.1. COMPREHENSIVE WHOLE-EXOME SEQUENCING ANALYSIS OF SPORADIC SCHIZOPHRENIA

Given the previous results, the enrichment of microtubule GO group comes as no surprise. It comprises elements from various already discussed groups like motor proteins, microtubule-based movement and cilium, and adds further gene groups like tubulins (TTL6, TUBB8).

Multicellular organismal homeostasis (GO:0048871)

Def.: *"Any process involved in the maintenance of an internal steady state at the level of the multicellular organism."*

Pvalue WebGesalt: 1.18×10^{-2} , **Pvalue** DNENRICH: 5.28×10^{-4}

The enrichment of genes of the multicellular organismal homeostasis seems to be, in contrast to most of the other categories, not specifically related to neuronal health. Although certain genes (*CNGB1*, *SCNN1A*) with functions in voltage-gated sodium and ion channels are included, there is, so far, no evidence for an involvement of this specific ontology group in the neuropathology of schizophrenia.

Actin-based cell projection (GO:0098858)

Def.: *"A cell projection supported by an assembly of actin filaments, and which lacks microtubules."*

Pvalue WebGesalt: 2.80×10^{-2} , **Pvalue** DNENRICH: 1.52×10^{-2}

Cell projection is an important mechanism for many biological processes like the cell motility, cancer-cell invasion, endocytosis, phagocytosis, exocytosis, pathogen infection, neurite extension and cytokinesis. Due to this wide range of components, I did not detect any clear association between the limited amount of categorized genes and neuronal processes suggesting that the enrichment might be a byproduct of the enrichment of related categories.

Taste receptor activity (GO:0008527)

Def.: *"Combining with soluble compounds to initiate a change in cell activity. These receptors are responsible for the sense of taste."*

Pvalue WebGesalt: 3.60×10^{-2} , **Pvalue** DNENRICH: 5.28×10^{-4}

The last category that reached statistical significance is the activity of taste receptors. So far, there is no evidence for any kind of associations between this receptor type and mental health. The enrichment clearly is most likely due to an accumulation of unrelated variants.

4.1.8.2 GO redundancy and genetic overlap

The hierarchical layout of GO terms consequently leads to a certain degree redundancy across multiple GO terms. As a result, a specific set of core genes can lead to an enrichment of multiple related terms without necessarily providing significant additional information. All terms that were affected by this redundancy to a high degree are shown in Table 4.1 along with the respective supersets.

ID	Name	Description	Pvalue	Superset
GO:0007018	Microtubule-based movement	<i>"A microtubule-based process that results in the movement of organelles, other microtubules, or other cellular components. Examples include motor-driven movement along microtubules and movement driven by polymerization or depolymerization of microtubules."</i>	WG: 5.39×10^{-5} DR: 1.70×10^{-4}	Motor activity
GO:0005875	Microtubule associated complex	<i>"Any multimeric complex connected to a microtubule."</i>	WG: 4.90×10^{-4} DR: 2.83×10^{-4}	Motor activity
GO:0044420	Extracellular matrix component	<i>"Any constituent part of the extracellular matrix, the structure lying external to one or more cells, which provides structural support for cells or tissues; may be completely external to the cell (as in animals) or be part of the cell (as often seen in plants)."</i>	WG: 2.37×10^{-3} DR: 3.72×10^{-2}	Extracellular matrix structural constituent
GO:0005578	Proteinaceous extracellular matrix	<i>"A structure lying external to one or more cells, which provides structural support, biochemical or biomechanical cues for cells or tissues."</i>	WG: 5.03×10^{-3} DR: 1.63×10^{-2}	Extracellular matrix structural constituent
GO:0005581	Collagen trimer	<i>"A protein complex consisting of three collagen chains assembled into a left-handed triple helix. These trimers typically assemble into higher order structures."</i>	WG: 2.80×10^{-2} DR: 2.35×10^{-2}	Extracellular matrix structural constituent

Table 4.1: Enriched GO terms with high redundancy. GO term descriptions according AmiGO2 [114]. The pvalue column specifies the enrichment pvalues after multiple-testing correction (WG: WebGestalt, DR: DNENRICH). Supersets indicate the GO set(s) that already covers most genes from the respective set.

4.1.8.3 Summary

Overall, the enrichment analysis clearly draws a picture of enriched microtubule and actin-related genes throughout all three GO categories (Biological process, Molecular function, Cellular component) as seven of 16 significant groups are directly related to one of those components. Motor activity genes play a crucial part due to the frequent occurrence of disrupted motor proteins dynein, kinesin and myosin that were found during multiple independent analysis steps. Additionally, several cytoskeletal (actin and microtubule) genes with clear associations to schizophrenia were found during manual gene inspection. As shown in section 3.3.2, almost 2/3 of all samples carry at least a single variation type within a gene of the neuronal cytoskeleton (microtubule, actin cytoskeleton, motor activity sets). Taken together, these findings suggest that the neuronal cytoskeleton might play an important role in the neuropathology of schizophrenia due to its involvement in various fundamental neurological mechanisms and pathways.

THE NEURONAL CYTOSKELETON - A COMMON FEATURE

As shown in the previous sections, I applied a unique combination of bioinformatic methods, selected to complement each other by probing differential aspects of the mutational spectrum. Throughout these analyses, the neuronal cytoskeleton can be identified as a common feature with an enrichment of various deleterious aberrations. In the following, I will provide more information about the biological background of the neuronal cytoskeleton and discuss and group my previous findings towards specific elements of the neuronal cytoskeleton.

5.1 Background: Neurodevelopment and the neuronal cytoskeleton

Neurodevelopment is a long-term process and consists of several steps. During the embryonic and adult neurogenesis, neuronal structures develop in multiple stages: cell proliferation, cell migration, and cell differentiation. During those steps, new neurons are generated that subsequently migrate towards their final destination. These newly generated cells then acquire their complex shape and specific connections by starting axon and dendrite outgrowth, establishing neuronal transmitter transport and by forming synapses for neurotransmission during synaptogenesis. These structures have to be maintained carefully to ensure the integrity of brain processes. A multitude of those neuronal processes rely on the neuronal cytoskeleton. It consists out of three main structures: microtubules, intermediate filaments (neurofilaments) and actin filaments (Figure 5.1).

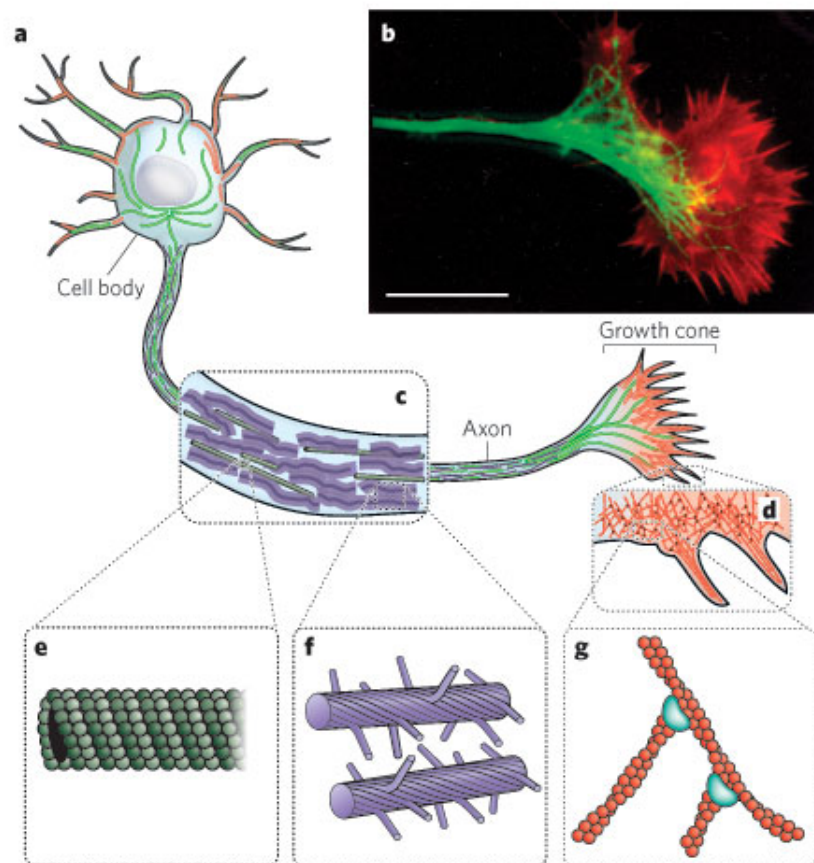


Figure 5.1: **(a,b)** A neuron with the neuronal cytoskeleton, consisting out of three main structures: microtubules (green), intermediate filaments (purple) and actin filaments (red). **(c)** The neuronal axon, a membrane-bounded extension, in which microtubule and intermediate filaments form a structural backbone that enables the transport of cargo (e.g. neurotransmitters) from the cell body to the synapse. **(d)** The growth cone, an actin-supported extension of a developing neurite. **(e)** A microtubule, formed like a hollow cylinder, consisting out of polymerised α - and β -tubulin dimers. **(f)** Neurofilaments provide structural support for the axon and regulate the axon diameter. **(g)** Actin filaments are arranged into networks and play a critical role in neuronal growth and secretion. Figure taken from [184]

5.1.1 The microtubule cytoskeleton, structural element and track for neurotransmitter traffic

Microtubules are one of the major cytoskeletal components of neurons. They are essential for many critical cellular and developmental processes like neuronal migration, polarity, and differentiation.

The base of the microtubule cytoskeleton is the microtubule-organizing center, also called centrosome. It provides the structural foundation for the microtubule array and also acts as the primary microtubule nucleation site. The centrosome plays a major role during the neuronal migration in which it stabilizes and directly regulates the movement of the neuron. Failures in maintaining the proper position of the centrosome during the migration can lead to incorrect placement of the neuron [185].

A well-known centrosomal protein previously associated with schizophrenia is disrupted-in-schizophrenia 1 (DISC1). DISC1 interacts with several centrosomal proteins and plays an important role in neuronal migration throughout the interaction with cytoplasmic dynein and the stabilization of nucleus-centrosome coupling [186]. Another centrosomal protein Pericentriolar material 1 (PCM1) which also has been associated with schizophrenia has important functions in localization of centrosomal proteins and microtubule anchoring indicating the severe effects of defects within the centrosomal system [185].

As microtubules have an intrinsic polarity, they also provide an ideal platform for directional and structured component transport which is required for development and maintenance of axonal and dendritic processes [187]. Motor proteins like kinesins and dyneins move along the microtubule structures and deliver their cargo to specific cellular regions [172]. Further processes in which microtubules are involved include the neurite initiation and outgrowth, the axon differentiation, elongation and regeneration as well as dendritic spine morphodynamics and synapse functioning [187].

Therefore microtubules have important functions not only during neurogenesis but also for neuron operation by ensuring cell stability and the reliable vesicle transport that is crucial for synapse activation.

5.1.2 Neurofilaments provide stability to mature axons

Neurofilaments are intermediate filaments found in neurons and are a major component of the axonal cytoskeleton. Compared to actin filament protein (16 nm diameter) and tubulin (6 nm diameter), they are an intermediate structure with a 10 nm diameter which explains the used terminology [188]. Together with other cytoskeletal structures, neurofilaments provide structural support for the axon and regulate the axon diameter [189]. Additionally, neurofilaments are also associated with axonal growth and maintenance of neuronal homeostasis [190]. Thus, it is not surprising that neurofilaments post-transcriptional gene regulation are assumed to play a role in neurodegenerative diseases [191].

5.1.3 Actin microfilaments play a critical role in dendritic spine modeling

Dendritic spines are tiny protrusions emerging from their parent dendrites. Dendrites of a single neuron can contain hundreds to thousands of spines which can be found in different morphological forms. Dendritic spines are able to store synaptic strength and help transmitting electrical signals to the cell body. Within those spines, a network of actin filaments is located that forms various kinds of higher order structures. These structures then ultimately determine the shape of the dendritic spine. Interestingly, it has been shown that dynamics of actin filaments are co-regulated with microtubule dynamics that might be essential for temporal and local regulation of dendritic spines. Actin dynamics are especially important since dendritic spines show plastic morphological changes depending on synaptic activity. It has also been noted that several neuropsychiatric, neurodevelopmental and neurodegenerative diseases show alterations in the morphology of dendritic spines [192].

5.2 The neuronal cytoskeleton and schizophrenia

As shown in the previous sections, the neuronal cytoskeleton is known to be an integral component for maintaining various neurological processes starting from embryonic and adult neurogenesis up to cell proliferation, cell migration, and cell differentiation. Therefore, variants affecting these pathways and processes can have severe effects on neurological functions and mental health. In my work, I find that the identified mutations particularly affect two main components of the neuronal cytoskeleton: microtubules and actin filaments. In terms of function, the affected genes can be assigned to certain groups, each responsible for specific neuronal processes like stability, transport or migration.

5.2.1 Microtubule stability, dynamics and the microtubule-organizing center

Microtubules are one of the major cytoskeletal components of neurons. They are essential for many critical cellular and developmental processes like neuronal migration, polarity, and differentiation.

The structural base of the microtubule cytoskeleton is the microtubule-organizing center, also called centrosome. The microtubule-organising center has critical functions in diverse neurodevelopmental processes. Here I identified mutations in several genes (*e.g. CCDC15, CCP110, CEP192, CEP350, CUL7, DLG1, TUBB8*) that were previously reported to be involved in this central organising element or directly affect microtubule stability and dynamics.

Especially centrosomal proteins like CEP192, which is expressed in multiple brain tissues, is known to play essential roles in microtubule organisation and stability throughout its interaction with NEDD1, AURKA and CYLD. [193].

Similar functions are accomplished by CUL7, whose depletion results in altered microtubule dynamics [153]. It is highly expressed in the cerebellar brain and furthermore plays a role in

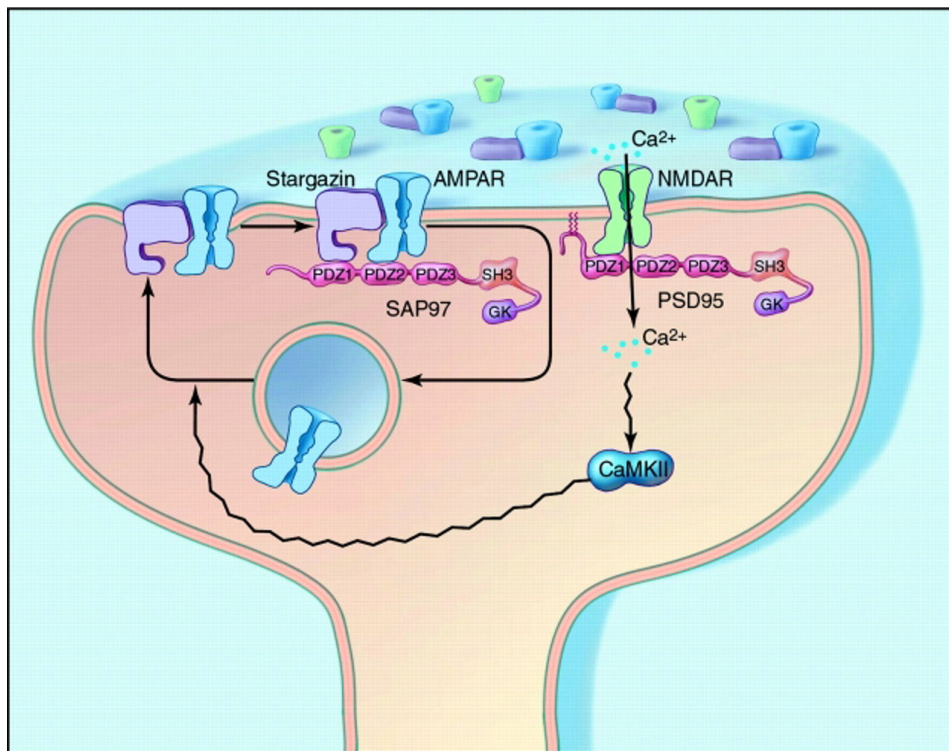


Figure 5.2: Regulation of synaptic glutamate receptors by PDZ proteins. PDZ proteins are characterised by a PSD-95/SAP-90, Discs-large, ZO-1 homologous domain (PDZ) motifs [194]. The synaptic targeting of AMPA receptors requires stargazin binding to PDZ proteins like SAP97/DLG1. In contrast to NMDA receptors, the expression of AMPA receptors is dynamically regulated depending on synapse activity. Figure taken from [194].

neural dendrite patterning and growth [154]. I also identified CUL7 as hub protein within the candidate PPI interaction network (Figure 3.13). Interestingly, another identified hub protein, Striatin (STRN), has also been shown to play a role in modulating microtubule dynamics and is assumed to play a role in dendritic Ca²⁺ signaling [152].

Another gene contributing to the dynamics of the microtubule network is *DLG1/SAP97* that is located at the schizophrenia-susceptibility locus 3q29 [195]. DLG1 modulates microtubules by dynein interactions [196] and has also been shown to interact with glutamate receptors (Figure 5.2) and form trafficking complexes that transport receptors along dendritic microtubules [197]. Additionally, Uezato et al. [195] found reduced cortical expression levels in patients with early-onset schizophrenia carrying a newly identified splicing variant of DLG1.

5.2.2 Microtubule-based transport

Whereas the microtubule-organizing center and microtubule stability and dynamics have an important role in maintaining the overall structure of the microtubule network, motor proteins are required for driving and organizing the transport of various cargo along the microtubule

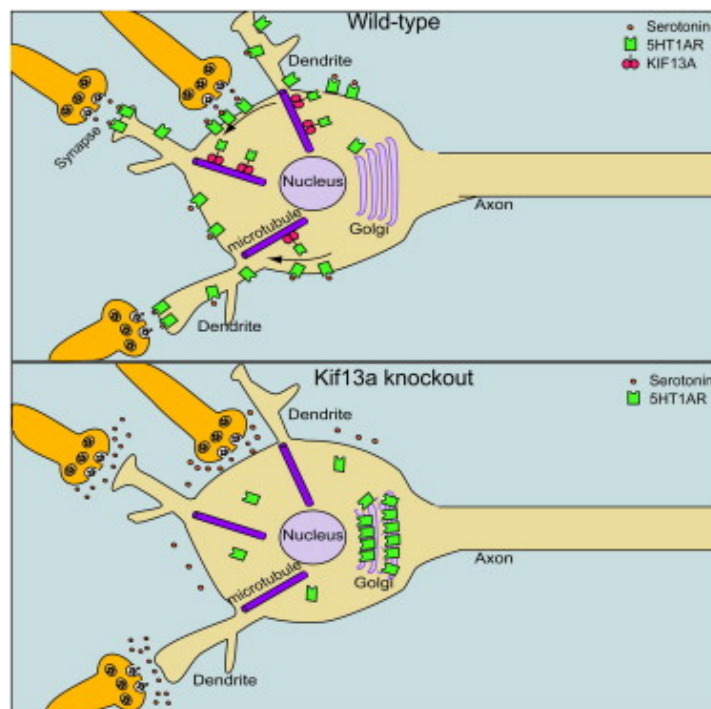


Figure 5.3: Model with Kif13a knockout. With functional 5HT1A receptor transport organised by Kif13a along microtubules, serotonin uptake is maintained at the dendrites. A knockout of Kif13a prevents this transportation of the receptors and consequently the transmission of signals. The released serotonin molecules remain in the synaptic cleft. Figure taken from [198]

structures. Here I found mutations in several kinesin (KIF13A, KIF13B, KIF1C, KIF20B) and axonemal and cytoplasmic dynein motor proteins as well as in the cytoskeletal linker protein dystonin that are all related to the microtubule-based transport.

Kinesins, like KIF13A, mediate the anterograde transport along the microtubules towards synapses. In the specific case of KIF13A, the transported cargo is assumed to be the Serotonin Type 1A Receptor (5HT1A), as a study of Zhou et al. [198] showed that 5HT1A receptors are not properly transported in Kif13a^{-/-} neurons using a mouse model (Figure 5.3). As a consequence, there is no serotonin uptake at the dendrites. Although I did not find a complete KIF13A-knockout, several patients accumulated deleterious variants across *KIF13A* which might result in disrupted 5HT1A receptor transport. Another example how mutations in motor proteins can affect crucial cellular processes, is a protein of the same family, KIF13B. The kinesin family member 13B is known as a molecular motor required for the trafficking of vascular endothelial growth factor receptor 2 (VEGFR2) that is involved in axonal wiring [199, 200] and therefore can play an essential role in brain development and maintenance.

In contrast to anterograde transport, the retrograde transport is mainly organized by cytoplasmic dynein. In this present study I did find only one disrupted cytoplasmic dynein (DYNC2H1) but nine axonemal dyneins which are mostly found in cilia and flagella. Although axonemal

dyneins are not directly associated with the microtubule-based transport like their cytoplasmic counterparts, a number of studies have shown that especially the primary cilia are critical for a number of brain processes, like the neural tube development, neuronal migration and differentiation [201]. Notably, also the well-established schizophrenia candidate gene *DISC1* is implicated in the formation and/or maintenance of the primary cilia as well as the targeting of dopamine receptors to the ciliary surface [202]. Still, I did find a *de novo* mutation within dystonin (*DST*). Dystonin is assumed to take part in the retrograde axonal transport [203]. A loss of function of this cytoskeletal linker protein has been shown to cause neurodegeneration in dystonia musculorum (*dt*) mutant mice [204] and was also previously linked to schizophrenia [205, 206].

5.2.3 Neuronal migration

Like transport, also neuronal migration critically relies on the stable underlying microtubule network in order to direct the movement of the neuron towards its final destination. Here, we identified three genes (*IFT74*, *KIF20B*, *RPGRIP1L*) that contribute to this essential neurodevelopmental process. *RPGRIP1L*, which is also located in the microtubule organizing center, has been associated with the coordination of migration and placement of neurons in the developing cerebral cortex as a knockdown of this and several other ciliopathy genes lead to a retarded neuronal migration [207]. Accordingly, *RPGRIP1L* was reported to be essential for normal brain development [208].

IFT74 (Intraflagellar Transport 74) is the second gene I identified to take part in neuronal migration. *IFT74* has been shown to be a critical factor in neuronal migration throughout its interaction with *PAX6* (Paired Box 6) [209]. *PAX6*, which is widely expressed in the central nervous system, regulates the regional development and neuronal migration in the cerebral cortex and was suspected to be involved in paranoid schizophrenia [209–211]. *IFT74* has therefore been identified as a regulatory element of neurogenesis [209]. The last gene I identified is another member of the kinesin motor protein family, *KIF20B*. Sapir et al. [212] showed that *KIF20B* promotes the polarization of migrating primary hippocampal and pyramidal neurons throughout its cargo *Shootin1* by using a knockout approach. Furthermore, it was concluded that *Shootin1/KIF20B* most likely act in the same genetic pathway [212].

5.2.4 Actin-based processes

Next to the microtubules, the actin cytoskeleton is required for neuronal processes like transport or the formation of dendrites and dendritic spines. Here, I identified several mutated genes involved in actin-related pathways that may lead to critical neuronal dysregulations or functional deficiencies.

Dendritic spines play an important role during neuronal signal transmission and neuron connectivity. Here I particularly found disruptions in *IQGAP2*, a gene that has been shown to be a key

regulator of dendritic spine number [213]. As a result, it was concluded that IQGAP1 contributes to the cognitive deficiencies in brain disorders [213]. Additionally, IQGAP1 is also an important component of N-methyl-D-aspartate receptor (NMDAR), a major glutamate receptor subtype. Disruptions of NMDAR function lead to altered neurodevelopment and therefore may be involved in progression and development of schizophrenia [214].

Besides, I identified another disrupted gene, *TRIOBP*, that, via its interaction with TRIO, indirectly regulates neural tissue development and controls actin cytoskeleton organization. It has also been shown that the 3' splice variant TRIOBP-1 aggregates in differentiated neurons of the brain in post mortem brain samples from schizophrenia patients and may directly affect cell development [174].

As for the microtubule-based transport, I also find disruptions in actin-based transport proteins, myosin IA, VIIB, IXA and VC, which are actin-based motor molecules with ATPase activity. Especially MYO1A, that is involved in directing the movement of organelles along actin filaments may play an important role. Rare mutations in *MYO1A* have been identified with autism [215] and it also directly interacts with the classical schizophrenia candidate gene disrupted in schizophrenia 1 (DISC1). Also MYO7B which has been associated with antipsychotic treatment response [216] and MYO9A that is assumed to regulate neuronal morphology and function, seem to be relevant targets for further investigations. Interestingly, heterozygous knockouts of MYO9A have been found to lead to decreased food intake (Table 3.3) in mice which can be a symptom of schizophrenia [119].

5.2.5 Cases with disrupted neuronal cytoskeleton

As shown in the previous sections, multiple findings indicate the enrichment of genetic disruptions within genes of the neuronal cytoskeleton. In order to determine the fraction of schizophrenia patients that might be explained by such genetic variants, I generated the sample-specific counts for affected genes of the neuronal cytoskeleton. As presented in Section 3.3.2, almost 2/3 of all samples carry at least a single damaging variation type within a gene of the neuronal cytoskeleton. Clearly, with the increasing number of affected genes, the chance of a significant impact on the phenotype rises. Based on the observed numbers (Figure 3.16), two main statements can be inferred.

Firstly, the point that 2/3 of all samples carry damaged neuronal cytoskeleton genes indicates that about 1/3 of the patients are unaffected. Clearly, for these individuals an alternative needs to be found in order to explain the phenotype. Although some causal mutations might be lost during data processing or simply follow a different and so far not analysed mechanism, one third of the samples seem to possess a genetically intact neuronal cytoskeleton. This observation supports the assumption about the importance of each individual genetic background, a fundamental basis of the personalised medicine.

Secondly, the other side of the sample spectrum contains multiple individuals with a high number

of cytoskeleton disruptions. Still, the interpretation is not straight forward. Eventhough the observed numbers indicate the existence of several individuals that might be explained by malfunctions of the neuronal cytoskeleton, a causal relation cannot be drawn by sequence data alone. As a result, it is not feasible to specify the exact samples that can be explained by the presented theory without further investigations.

Another noteworthy aspect that can be considered is the age of schizophrenia onset in cases with disruptions in the neuronal cytoskeleton. Based on the available data of the included trio patients (Figure 3.5) and literature [217], schizophrenia typically emerges in late adolescence and early adulthood. Therefore one could ask the question whether these disruptions only have an impact after adolescence. While this could potentially be the case, the crucial roles of the neuronal cytoskeleton in various neurodevelopmental processes such as neuronal migration and neurogenesis, an earlier impact of these variants seems plausible. As shown in Figure 1.2, early disturbances in the neuronal development can accumulate over time and might lead to a psychotic phenotype after adolescence. Such a model could also be assumed for schizophrenia cases with disruptions in the neuronal cytoskeleton.

TECHNICAL DISCUSSION

Next to the biological interpretation and discussion of the results presented, several technical issues exist that are essential to understand and assess the validity of the findings. In the following sections, I will outline multiple important technical aspects that influenced the outcome of this study.

6.1 Data quality

As described in Section 3.1 and shown in Figures 3.1 and 3.2, the overall quality of the used sequencing data met the requirements that allow an appropriate analysis of NGS data. Although slight differences in terms of target region coverage between the two data subsets (Eurofins and TUM Weihenstephan) were detected, both sets could be combined into a single comprehensive data set. Overall, all investigated quality and statistical parameters (e.g. base quality, mapping quality, amount of total variation) matched the expectations for standard high-quality Illumina sequencing data.

6.2 PCR duplicates - PCR effect or reading same the fragment twice

PCR duplicates are a common problem and widely discussed issue in NGS data analysis. PCR duplicates are typically defined as sequence reads that result from sequencing two or more copies of the exact same DNA fragment. While in the best case, this will only result in an amplification of the correct genomic DNA sequence, it can also contain erroneous mutations introduced during PCR amplification. While on the one hand, leaving such PCR duplicates untouched can introduce

false positive variants, PCR duplicate removal can lead to a large loss of data and possible true positive variants. Here, I will briefly present both positions with regard to current literature and discuss my decision not to remove PCR duplicates during the analysis of this dataset.

Typically, sequence reads are tagged as PCR duplicates whenever they have the same exact start position in the genome with small differences depending on the applied software. For example, Picards MarkDuplicates identifies read pairs that have the exact same 5' start position in the mapping as duplicates. In this case, the read pair with the highest sum of base qualities with $Q \geq 15$ will be kept. All other reads will be marked as duplicates which removes them from all further analysis steps.

The point why one would want to remove such artefacts is mostly motivated by false positive variant calls. If a mutation is introduced during one of the early steps of read amplification, this variant will occur in multiple reads, most likely leading to heterozygous or even homozygous variant call. To avoid this, PCR duplicates have to be removed. The obvious advantage is that the remaining variant calls have a fewer chance of being sequencing errors.

Still, it is clear that two reads that have identical mapping positions do not necessarily represent the same underlying cDNA fragment. Reads might have internal variation that indicate e.g. alternative isoforms or are simply due to different maternal and paternal alleles of a diploid genome. Removing duplicates purely on the basis of mapping coordinates has the disadvantage that it is not possible to distinguish between such cases. But depending on the data set, one will lose about 10 % of the actual sequence data.

Interestingly, a recent publication by Ebbert et al. [218] compared variant datasets when using Picard MarkDuplicates for duplicate removal, SAMTools for duplicate removal, or leaving duplicates untouched. The authors showed that there is no significant effect on the accuracy of subsequent variant datasets, raising the question whether PCR duplicate removal is truly necessary or not.

Based on the information presented, I decided to not remove PCR duplicates during this analysis. Although this might introduce false variant calls in the raw set of mutations, these artefacts can still be removed during later stages of filtering. Furthermore, by this approach, no true data will be lost due to wrong assumptions.

6.3 Phasing strategy and its impact on variant calling

Genotype phasing is a crucial part of every analysis that tries to detect complex variants like *de novo* or compound heterozygous variants. In both cases, knowledge about the underlying haplotype blocks is necessary to accurately determine these variants (Figure 6.1).

Depending on the available data, there are multiple ways to achieve this goal. While for complete family datasets (e.g. trio or quad studies) the presence of the parental genomes can be used for deciphering the correct genotype phase, individual genomes can be phased by using read-backed

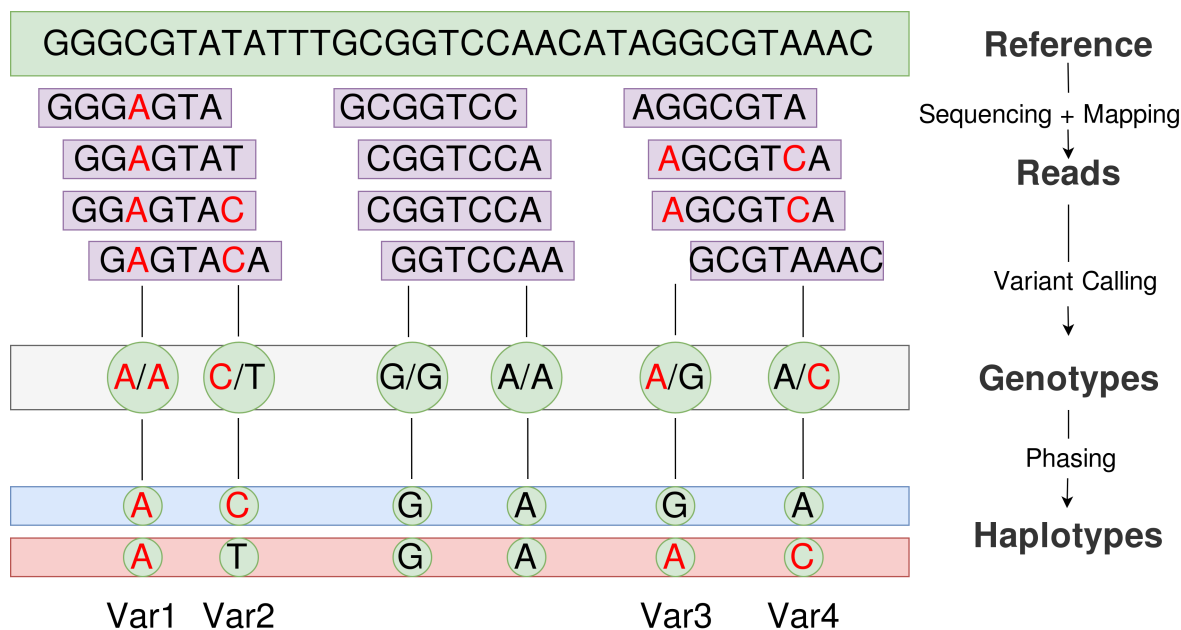


Figure 6.1: Schematic representation of haplotype phasing. By aligning reads against the reference genome, a set of genotypes can be determined. Throughout phasing, the respective haplotypes are extracted. Now, it is possible to call more complex variants like compound heterozygous variants (e.g. Var2 + Var3 and Var2 + Var4)

and reference-panel approaches. In my study, I applied all of these methods to obtain phased genotypes for all available samples including trio and individual samples. In the following, I will briefly discuss these strategies and how they influence downstream analyses.

The easiest and best scenario is the family-phasing, in which one has access to all relevant genomes (e.g. patient and parental genomes). In this case, the genotype phase of the subject can often directly be inferred by looking at the parental sites. This typically leads to the most accurate phasing results. In addition, this type of phasing is independent of genotype frequencies. A typical scenario in which this type of phasing is not able to determine the phase are so-called 'triple het sites'. In this case, all three samples have heterozygous genotypes which makes it impossible to determine the genotype phase based on this information alone.

For individual subjects, a more sophisticated approach is required due to the absence of supporting parental genomes. In order to cope with this situation, two potential information sources can be used: the actual read data or reference panels.

Physical phasing using sequence read data allows to determine whether certain variants are within the same haplotype block. This information is obtained when variants are observed within a single read fragment (Figure 6.1 (Var3 and Var4)). This approach is feasible for variants that are located close to each other but gets more and more complex with increasing distance between variants. In this case, multiple additional variants with supporting read data are necessary to provide a direct connection between two variants of interest. Depending on sequence depth and

Strategy	Target	Requirements
Family-based Phasing	All variant types	Only available in family studies
Read-based Phasing	All variant types	Requires high sequence depth and long read length
Reference Panel Phasing	Mostly common variants	Not well suited for rare variants

Table 6.1: Overview about the three different phasing strategies that were used during this work, as well as their requirements and primary phasing targets.

read length this often limits the number of variants than can be phased.

An alternative approach to this challenge is the usage of phased reference panels. Population-based phasing methods work by pooling linkage disequilibrium information across individuals. In contrast to read-based methods, they are independent of the actual read data but mostly rely on common variants that occur with a high frequency within the respective population. For the variants included, haplotype information is then inferred from the overall population providing an accurate mechanism for improved phasing accuracy.

Conclusively, all three methods that were applied in the course of this study (Section 2.2.4) can provide access to reliable phasing information but also come with certain drawbacks that directly influence downstream analyses (Table 6.1).

6.4 The influence of sample size on research conclusions

Schizophrenia sequencing studies in recent years by Purcell et al. [57] and Sekar et al. [219] have shown that often huge datasets comprising several thousand individuals are used when investigating schizophrenia genetics. Based on this fact, the question arises whether sequencing studies lacking this amount of data are expected to produce reliable results.

When discussing this question, it is inevitable to consider the genetic background of schizophrenia. As shown in Section 1.3.5, schizophrenia is assumed to be, for a large part, caused by various genetic variants that, in combination, explain this complex phenotype.

As a result, a larger sequencing cohort will consequently increase the probability of finding a significant fraction of the underlying rare functional variants. Still, also medium-sized studies, comparable to the presented one, should be able to detect ultra rare variants if a causal connection to the phenotype exists. If a causal relation between a certain variant and the phenotype exists, a clear bias towards this genetic mutation should be detectable.

Another aspect to consider is the definition of schizophrenia itself. As stated earlier, schizophrenia is rather a spectrum disorder, a composition of multiple phenotypes, instead of a clearly defined and distinct disease. Due to this characteristics, a cohort of several hundred individuals might in

fact contain multiple genetic subgroups depending on the study design. Therefore, even a smaller but well selected group of phenotypic-homogeneous individuals can increase the overall chance of detecting causal variants specific to this genetic subtype. This also holds true for ethnic groups that are often mixed during larger meta-studies. Although the integration of multiple datasets focusing on specific populations could increase detection power, it also carries the risk of diluting genetic signals.

6.5 Association vs. Causation

Association does not imply causation. This undeniable fact is a decisive factor in the evaluation of research results, as these two terms can easily be confused. An example for the relationship between both terms is the observation that people who daily drink more than 4 cups of coffee have a decreased chance of developing skin cancer. Obviously, this does not necessarily imply that coffee confers resistance to cancer. Instead, it could be assumed that people who drink a lot of coffee work indoors for long hours and thus have little exposure to the sun. Since increased sun exposure is known to increase skin cancer risk, the reduced amount of radiation leads to an decreased chance of developing skin cancer. Consequently, the initial observation is insufficient for inferring a direct causal link between coffee consumption and skin cancer. It merely suggest a common cause or some kind of direct or indirect relationship. Thus, association does not imply causation [220]. Likewise, the observed association of a given variant and a specific trait does not necessarily indicate a causal relation. The observed variation might be in a linkage disequilibrium with various other variants that explain the detected signal.

As a result, the aim of a research project that investigates the genetic causes of any disease clearly should be to determine the exact genetic factors that contribute to the overall disease risk. Although for monogenetic diseases this goal appears feasible, diseases affecting mental health are often far more complex. Therefore, genetic associations are often the primary starting point for each comprehensive analysis. To avoid misleading associations that are only side-effects of the causal connection, single associated targets must be placed in an overall context. This can be achieved by a multitude of methods depending on the actual research aims.

Clearly, the genetics of schizophrenia are far more complex compared to monogenetic diseases. True causal connections are, and will be hard to detect. In this study I used several methods (Sections 3.3, 3.3.1, 3.3.2 and 5.2) to set the found associations into context of biology and previous studies. Still, without experimental validation the proposed findings will need further investigation to confirm their relevance and correctness. Nevertheless, it is an important step that leads the way from raw associations into the direction of true causation.

6.6 Candidate genes in schizophrenia

When investigating genetic diseases, candidate genes or variants are often the result of several studies. They represent the core findings of many research papers and are often used as a starting point for follow-up analyses. For many mono-genetic diseases that are caused by a single gene or just few variants, this approach helps to identify putative targets which then easily can be experimentally verified. However in case of complex disease (e.g. schizophrenia) in which many factors contribute to the disease risk, a candidate gene approach can often result in numerous potential targets. In addition, most research projects are based on private data and use in-house or custom pipelines results which introduces a second degree of variance. As a consequence, three major problems arise: the lack of reproducibility, the constant increase of putative candidates and the specific characteristics of associated variants/genes.

In case of schizophrenia the lack of reproducibility has been described for 25 well-known historical candidate loci that were assumed to have strong associations with the disease by two recent studies [221, 222]. In both cases, the authors concluded that there is no evidence that schizophrenia candidate genes are more associated with schizophrenia than non-candidate genes. As a result, many candidate genes that appear to contribute to the genetic risk of the schizophrenia spectrum could turn out as false positives.

The second issue is related to the complexity of the disease. Due the assumption that probably more than 100 genes are potential schizophrenia risk factors, the number of proposed candidate genes quickly rises. A recently established database for schizophrenia genes SZGR 2.0 [223] collected candidates from multiple sources by systematic review and curation of multiple lines of evidence. At the time of publication, the SZGR database included about 4200 common variants reported in genome-wide association studies, more than 1000 *de novo* mutations discovered by large-scale sequencing of family samples, 215 genes spanning rare and replication copy number variations, 99 genes overlapping with linkage regions, 240 differentially expressed genes, 4651 differentially methylated genes and 49 genes as antipsychotic drug targets [223].

Although the SZGR database is just one source of candidate genes, this provides a good impression about the current state of candidate genes in schizophrenia research. Each year, new putative candidates are found, further enlarging the pool of schizophrenia-associated genes. An obvious problem of this development is that overall group of candidates includes many false positive signals making it difficult for researches to discriminate between a potential true candidate and a more likely false positive.

The last issue are the special characteristics of schizophrenia-associated variants and genes. Due to the multifactorial nature of the disease, causal variations that contribute to the disease might appear in non-affected individuals as well. Assuming a model in which a certain threshold of disease-contributing variants has to be exceeded to cause the phenotype, such variants would also be present in the general healthy population. This is in strong contrast to monogenetic diseases, where it is generally assumed that a single pathogenic variant occurs only in affected

individuals. As a result, the detection of candidate genes becomes even more difficult.

CONCLUSION AND OUTLOOK

The term 'schizophrenia' describes a mental illness that can affect many aspect of life. Worldwide, more than 21 million people suffer from its severe consequences. Numerous research studies are conducted each year, providing new insights into this medical phenomenon. Throughout the help of twin and family studies, it was shown that schizophrenia, for a large part, is determined by human genetics [4]. However, the specific genetic factors that cause the disease are still subject of discussion.

Due to the steady progress of DNA sequencing technologies that have been established during recent years, it has now become possible to analyse human genetics in large scale. These advances enable researchers to uncover previously unknown genetic links and their connections to schizophrenia. Within the course of this dissertation, I examined the genetics of well-selected schizophrenia trios and additional individual patients in order to gain novel insights into this complex disease. By applying a unique set of bioinformatic methods that cover the investigation of a broad range of genetic variant types, I made use of state-of-the-art technology to establish novel links between genes and the schizophrenia phenotype.

7.1 Schizophrenia and the neuronal cytoskeleton

Throughout my thesis, I was able to analyse the full set of provided patient data in a comprehensive and profound way. By doing so, I compiled a set of candidate genes, each related to schizophrenia by specific links, build by various applied methods and tools. This included the analysis of *de novo* variants, compound heterozygous and LoF mutations as well as more general investigations like the Edgetics and Genetic Burden approach.

Overall, I found 468 of such candidates. To make sense out of this comprehensive list and to reduce the effect of false positive findings, I projected the raw list of genes into the broader context

of genetic networks and pathways. By doing so, I was able to reveal an over-representation of mutated genes within the neuronal cytoskeleton which is essential for a multitude of neurological processes like neurotransmitter transport, neuronal migration and the formation of dendrites and dendritic spines. These results suggest a potential link between malfunctions of the neuronal cytoskeleton and schizophrenia.

7.2 Novelty of the work and contribution to the field of schizophrenia

Through the decades, schizophrenia research has led to various hypotheses about its etiology and pathophysiology, each trying to explain the processes related to the disease. While some of these might remain pure hypothesis, several tend to complement each other, contributing to solving the puzzle of schizophrenia. In this work, I presented a broader picture of schizophrenia and its connections to the neuronal cytoskeleton. Although the idea of the neuronal cytoskeleton as a potential therapeutical target in neurodegenerative diseases and schizophrenia was published earlier [224], so far, no comprehensive analysis provided further evidence for this connection. Throughout the multi-layer approach, combining various methods and tools, I provided new insights and a potential hypothesis for explaining certain aspects of schizophrenia. Due to the essential functions of the neuronal cytoskeleton in numerous neuronal pathways and processes including neurotransmission, the hypothesis fits well to previously published models like the dopamine hypothesis of schizophrenia that strongly rely on neurotransmission malfunctions (See Section 1.3.3.3). A study performed by Gabriel et al. also showed that dopamine transporter recycling relies on a dynamin-dependent mechanism that acts in concert with the actin cytoskeleton [225] which further supports the close relation between both components. More recent findings that focus on specific gene sets like the NMDA receptors or the activity-regulated cytoskeleton-associated protein (ARC) can also be brought into context with the neuronal cytoskeleton [226, 227].

Nevertheless, it has to be taken into account, that the present study was performed on a limited amount of samples. In order to validate the presented findings, it will be necessary to incorporate large sample sets, similar to the present large scale studies of Sekar et al., Purcell et al. or Ripke et al. [57, 112, 219].

7.3 Limitations and future directions

All studies have limitations. In the following, I conclude on the most important ones that affected this study and will provide recommendations and ideas for further work and follow-up research.

Phenotyping

The number of samples included in a study is an obvious, but sometimes misleading criterium in order to assess the scientific value of a certain study. While large studies clearly improve the overall statistical power and the probability to detect rare variants, it is inevitable to maintain and improve patient phenotyping. Especially for disorders like schizophrenia, which rather describe a spectrum of symptoms and includes a collection of phenotypes summarised under the common term, accurate phenotyping is a crucial key for improving our understanding of complex mental diseases. This is also important due to the lack of biomarkers that could improve diagnosis and stratification of schizophrenia patients.

Whole-exome vs. whole-genome sequencing

Due to the recent and foreseeable advances in [NGS](#) technologies, whole-genome sequencing has and will increasingly become the way to go. While exome sequencing offered a cheaper possibility to analyse the most relevant genomic parts, it comes with certain limitations. Three prominent constraints are the inability to call genome-wide Copy Number Variations ([CNVs](#)), the effect of exome capturing kits and the lack of sequencing data for non-coding regions.

The importance of [CNVs](#) has recently been shown by a large scale genome-wide study, incorporating 41,321 subjects that detected eight loci with genome-wide significant evidence [[228](#)]. Also previous studies strongly suggested the relevance of [CNVs](#) in the genetic etiology of schizophrenia [[228](#)]. Due to the obvious methodological restrictions of exome sequencing, which focuses on only selected exonic regions, a comprehensive [CNV](#) analysis is hardly feasible. Due to this restriction and the undisputed implication of [CNVs](#), whole-genome sequencing should be favoured of whole-exome sequencing whenever possible.

A second limitation that comes into play when analysing whole-exome data, is the effect of exome capturing kits. These kits, predetermine the exact area in which potential genetic variants and associations can be found. Although exome kits typically include most genes, the technical targeting for specific genomic regions leads to the loss of certain areas in which target capturing failed.

The third issue is the lack of sequencing data for non-coding regions of the genome. As highlighted by Gloss and Dinger [[229](#)], the impact of variation across the complete genome is required to realise the full potential of clinical genomics. This is a crucial aspect since non-coding regions contain valuable information about structural, regulatory, and transcribed regions [[229](#)]. Although the interpretation of non-coding variants is still challenging, new approaches, such as deep learning, were already successful in assessing the impact of various non-coding variants [[230](#)].

In conclusion, all three limitations are easily avoidable when choosing a whole genome approach which should therefore be used whenever feasible.

Experimental validation of sequencing variants

Systematic evaluation of **NGS** sequencing technologies has shown that the typical error rate of sequencing-by-synthesis technologies like the Illumina HiSeq systems is around 0.1% per nucleotide [231] providing a high standard for nucleotide sequencing. Still, errors occur and without proper quality controls, mis-called nucleotides can quickly become an issue.

In addition, the processing of **NGS**-data is error-prone itself. Throughout complex steps like genomic mapping, realignment or base quality score recalibration (See Section 2), errors can quickly arise. Taken together, those two issues inevitably will lead to a number of false positive results, that, depending on the actual research goals, can critically influence the outcome. Especially in case where multiple genomes are involved (e.g. trio/ *de novo* studies), a single sequencing error within one of the related genomes will lead to false-positive/false-negative *de novo* variant calls. As a result, experimental validation should be applied whenever feasible. For doing so, Sanger sequencing is probably the most common method used. With read-lengths of up to 1000 bp and per-base accuracies as high as 99.999% [232], it provides an accurate way to validate specific regions of interest.

Clearly, the lack of such experimental validation is a limitation for this study, leaving the possibility of several false positive findings. Although careful filtration and thoughtful analysis can reduce the amount of such cases, an experimental validation is recommended for any follow-up study.

Animal models and pluripotent cells

While validation of specific sequences provides information about the exact genomic picture, it cannot validate detected associations between certain **SNPs** and a specific phenotypic trait. In many studies, this is handled by animal models. These extensively studied non-human species, e.g. mouse or zebrafish, can be used to monitor the effect of specific genomic alterations that are artificially introduced.

Still, creating animal models for schizophrenia can be challenging. Due to the complex nature of the disease, single schizophrenia-associated variants/genes are not expected to be sufficient for causing a clear schizophrenia phenotype. Therefore, even if variants are correctly mimicked, the effect of a single variant will most likely be limited. Still it should be mentioned that mouse models for abnormal nervous system physiology, behaviour and related phenotypes have been successfully established.

Recent advances in human stem cell research have now enabled new possibilities for disease modelling that, in future, could provide a reliable way for disease simulation and genetic validation. This approach is mainly based on the ability to induce pluripotency [233], differentiation [234] and the achievement of accurate and efficient genome editing [235].

In 2006, Takahashi et al. showed that mouse pluripotent stem cells can be directly generated from fibroblast cultures by the addition of four genetic factors (Oct3/4, Sox2, c-Myc, and Klf4) [236].

A year later, pluripotent stem cell lines were already derived from human somatic cells [233]. Throughout this artificial conversion, these cells gain the possibility to differentiate into specific cell types. In 2010, Vierbuchen et al. successfully converted mouse embryonic and postnatal fibroblasts into functional neurons [237]. A year later the induction of human neuronal cells was achieved [234]. These groundbreaking advances not only provide the ability to artificially generate specific cell types but even could be used to generate and study brain organoids [238]. The second essential part of this idea, is the accurate and efficient genome editing method, known as the CRISPR/Cas system [235]. This method, based on the CRISPR/Cas protein complex, allows to accurately re-program DNA sequences. This enables the introduction of specific variations like SNPs and Indels or even complete genetic knockouts.

Taken together, those three methods (Induction of pluripotency, cell differentiation and the CRISPR/Cas system) allow to accurately define a specific genetic cellular background that can be introduced into artificial human neurons. Although the mentioned technologies are new and need further improvements, it already offers a powerful application across basic science, biotechnology, and medicine.

7.4 Conclusion

Undeniably, schizophrenia is an extensively studied disease. Throughout decades of research, scientists have tried to shed light onto the dark matter of schizophrenia. Despite this tremendous effort, the neural basis of mental diseases like schizophrenia is one of the 23 yet unsolved problems in neuroscience [2].

Recent technological advances in DNA sequencing have made it possible to analyse the genetic background of schizophrenia at large scale. In my study, I aimed to detect novel genetic findings in a carefully selected cohort of schizophrenia patients. I could achieve this goal by combining various computational and system biological methods. The presented results indicate a strong role of the neuronal cytoskeleton, that underlies numerous neurological processes and pathways. I detected various damaging mutations across a large number of patients in multiple essential parts of this neuronal system. These candidate genes provide a profound starting point for further follow-up studies.

I was able to show the concordance of my findings to previous studies made by independent researches as well as the novelty of my work within the field of schizophrenia sciences.

With the coming of new technologies like the CRISPR/Cas system and the advances in human stem cell research, it might soon be possible to completely grasp the genetics of schizophrenia. This could ultimately resolve the mystery of schizophrenia, and as a next step, provide a reliable way in order to find successful treatments for more than 21 million people worldwide.

APPENDIX

This section lists the genetic findings for each individual schizophrenia patient that was included in the analysis and for whom genetic variants were found. In each case, the number of detected variants/genes is specified. A description of the used abbreviations can be found at the end of the table.

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ103	-	-	4	-	1	0	0	
SZ104	-	-	2	-	1	1xSZ	1	
SZ105	-	-	1	-	0	1xCSZ, 1xSSZ	0	
SZ107	-	-	2	-	3	0	3	
SZ108	-	-	1	-	0	1xCSZ	2	
SZ109	-	-	1	-	0	0	2	
SZ111	-	-	4	-	1	2xSZ	2	
SZ112	-	-	3	-	0	1xSZ, 1xCSZ	1	
SZ113	-	-	0	-	0	1xCSZ	0	
SZ114	-	-	3	-	0	1xSZ, 4xCSZ	3	
SZ115	-	-	3	-	0	2xCSZ	1	
SZ117	9	0	6	0	1	0	4	*
SZ121	-	-	0	-	0	1xCSZ	3	
SZ123	-	-	1	-	1	0	0	
SZ124	-	-	1	-	0	0	0	
SZ125	-	-	3	-	0	0	1	
SZ126	-	-	0	-	0	1xCSZ	0	
SZ127	-	-	3	-	0	2xCSZ	0	
SZ128	-	-	2	-	0	1xCSZ	0	
SZ129	-	-	1	-	1	1xCSZ	1	
SZ130	-	-	3	-	0	1xSZ, 1xCSZ	1	
SZ131	8	0	7	1	0	1xCSZ	0	*
SZ132	-	-	1	-	2	0	0	

CHAPTER 8. APPENDIX

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ133	-	-	3	-	0	4xCSZ	2	
SZ135	-	-	2	-	0	1xCSZ	0	
SZ136	-	-	2	-	0	1xSZ	1	
SZ137	-	-	1	-	1	0	2	
SZ139	-	-	0	-	0	1xSZ, 2xCSZ	1	
SZ140	-	-	2	-	0	0	1	
SZ141	-	-	1	-	1	1xSSZ	1	
SZ143	-	-	3	-	1	1xSZ	0	
SZ144	-	-	3	-	1	0	0	
SZ146	-	-	0	-	1	1xCSZ	0	
SZ147	11	0	3	5	1	1xCSZ	1	*
SZ148	-	-	1	-	1	3xSZ	0	
SZ149	-	-	1	-	0	0	1	
SZ150	-	-	1	-	1	1xCSZ, 1xSSZ	0	
SZ151	-	-	4	-	2	0	2	
SZ152	-	-	4	-	0	1xCSZ	2	
SZ153	-	-	1	-	0	1xSSZ	1	
SZ154	13	0	4	1	1	3xCSZ	2	*
SZ158	9	1	5	0	1	0	2	*
SZ159	-	-	1	-	0	2xCSZ	1	
SZ160	-	-	2	-	0	0	0	
SZ161	-	-	1	-	1	0	0	
SZ162	-	-	2	-	1	1xSZ, 1xCSZ	0	
SZ163	-	-	1	-	1	1xCSZ	1	
SZ164	-	-	1	-	1	0	0	
SZ166	-	-	1	-	0	0	1	
SZ167	-	-	2	-	0	1xSZ	1	
SZ168	-	-	2	-	0	1xCSZ, 1xSSZ	1	
SZ170	-	-	3	-	0	0	1	
SZ171	13	1	6	2	0	0	1	*
SZ172	-	-	3	-	1	1xSZ	0	
SZ174	-	-	1	-	1	0	1	
SZ176	-	-	2	-	0	0	1	
SZ177	-	-	3	-	0	1xCSZ	1	
SZ178	-	-	2	-	1	0	1	
SZ179	-	-	1	-	1	1xCSZ	2	
SZ181	12	0	7	1	1	0	2	*
SZ182	-	-	2	-	0	0	0	
SZ183	-	-	2	-	0	0	2	
SZ184	-	-	2	-	0	0	2	
SZ185	10	0	3	3	0	0	1	*
SZ186	-	-	0	-	0	1xCSZ	3	
SZ187	-	-	3	-	0	1xCSZ	2	
SZ190	-	-	2	-	1	0	1	
SZ192	-	-	0	-	0	1xCSZ	2	

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ193	-	-	3	-	1	0	2	
SZ194	-	-	4	-	0	0	1	
SZ195	-	-	3	-	0	1xSZ	2	
SZ196	-	-	7	-	0	1xCSZ	3	
SZ197	-	-	6	-	0	1xCSZ	2	
SZ198	-	-	2	-	0	1xCSZ	2	
SZ199	-	-	1	-	0	0	0	
SZ200	9	0	2	3	0	0	3	*
SZ202	-	-	1	-	1	0	1	
SZ203	-	-	2	-	0	2xSZ, 2xCSZ	0	
SZ204	-	-	2	-	0	1xCSZ	2	
SZ207	-	-	2	-	0	0	1	
SZ208	-	-	4	-	1	1xCSZ	1	
SZ209	-	-	4	-	0	1xCSZ	1	
SZ210	-	-	3	-	0	2xSZ	0	
SZ211	-	-	4	-	0	1xCSZ	0	
SZ212	-	-	1	-	1	0	1	
SZ213	-	-	0	-	0	0	1	
SZ216	14	0	1	2	0	0	3	*
SZ220	-	-	1	-	0	1xCSZ	1	
SZ221	-	-	3	-	1	1xCSZ	0	
SZ223	-	-	4	-	1	1xCSZ	1	
SZ225	-	-	2	-	0	1xCSZ	0	
SZ226	-	-	2	-	0	0	1	
SZ227	-	-	4	-	1	1xCSZ	1	
SZ229	-	-	1	-	0	0	1	
SZ231	-	-	2	-	0	1xCSZ	0	
SZ232	-	-	1	-	0	0	1	
SZ233	-	-	2	-	0	1xCSZ	2	
SZ235	-	-	2	-	0	0	2	
SZ236	-	-	2	-	1	1xCSZ	1	
SZ237	-	-	2	-	0	1xCSZ	1	
SZ238	15	0	1	5	0	2xCSZ	3	*
SZ239	-	-	1	-	0	0	0	
SZ241	-	-	2	-	1	0	0	
SZ242	-	-	1	-	0	0	1	
SZ243	-	-	1	-	1	1xSZ	0	
SZ244	-	-	2	-	0	0	4	
SZ246	-	-	2	-	1	2xCSZ	1	
SZ248	-	-	4	-	0	2xSZ, 1xCSZ	2	
SZ249	12	0	1	1	2	0	2	*
SZ250	-	-	2	-	0	2xSZ	1	
SZ251	-	-	3	-	1	0	0	
SZ252	-	-	0	-	0	0	3	
SZ255	-	-	1	-	0	1xSZ	2	

CHAPTER 8. APPENDIX

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ256	-	-	2	-	1	0	0	
SZ259	-	-	1	-	2	1xSSZ	0	
SZ260	-	-	1	-	0	2xSZ	1	
SZ261	-	-	1	-	0	0	0	
SZ262	-	-	3	-	0	0	1	
SZ265	-	-	2	-	1	1xCSZ	1	
SZ266	-	-	2	-	0	0	1	
SZ267	-	-	2	-	2	1xCSZ	1	
SZ268	-	-	4	-	0	0	2	
SZ270	14	1	3	1	1	2xSZ, 3xCSZ	0	*
SZ271	-	-	0	-	2	1xSZ, 2xCSZ	1	
SZ272	-	-	2	-	1	0	1	
SZ273	-	-	2	-	0	0	2	
SZ274	-	-	0	-	0	1xCSZ	3	
SZ276	16	0	7	1	0	1xSZ	1	*
SZ277	14	1	3	0	0	1xSZ	3	*
SZ278	-	-	1	-	0	2xCSZ	2	
SZ279	-	-	3	-	0	1xSZ	0	
SZ280	-	-	2	-	0	0	1	
SZ281	15	2	3	0	0	0	2	*
SZ283	-	-	1	-	0	1xSZ	4	
SZ289	11	0	2	0	0	1xCSZ	1	*
SZ290	-	-	4	-	2	1xSZ	2	
SZ291	-	-	3	-	0	0	3	
SZ292	-	-	5	-	0	0	2	
SZ293	-	-	3	-	0	4xSZ, 2xCSZ	2	
SZ294	-	-	2	-	0	1xCSZ	2	
SZ296	-	-	3	-	0	1xCSZ	0	
SZ297	-	-	0	-	1	0	1	
SZ298	-	-	1	-	0	0	0	
SZ299	6	0	8	1	1	0	2	*
SZ300	-	-	1	-	1	0	2	
SZ302	12	0	2	0	0	0	2	*
SZ303	-	-	1	-	0	1xCSZ	0	
SZ304	-	-	1	-	0	0	0	
SZ305	-	-	2	-	0	0	0	
SZ306	-	-	3	-	1	0	3	
SZ308	-	-	6	-	0	1xCSZ	1	
SZ309	12	1	5	1	0	0	3	*
SZ312	-	-	2	-	0	0	0	
SZ313	-	-	0	-	2	1xCSZ	2	
SZ315	-	-	2	-	0	1xSZ	1	
SZ317	-	-	2	-	0	0	1	
SZ321	-	-	2	-	0	2xCSZ	1	
SZ322	-	-	1	-	0	1xCSZ, 1xSSZ	1	

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ324	-	-	9	-	0	1xSZ, 2xCSZ	1	
SZ325	-	-	2	-	0	1xCSZ	3	
SZ326	-	-	1	-	1	0	0	
SZ327	-	-	1	-	0	0	0	
SZ328	-	-	4	-	0	0	2	
SZ329	-	-	2	-	0	1xCSZ	0	
SZ330	-	-	3	-	1	0	1	
SZ331	9	0	2	2	0	1xSSZ	1	*
SZ333	-	-	1	-	1	0	0	
SZ336	13	1	4	3	0	0	2	*
SZ339	-	-	4	-	0	1xCSZ	1	
SZ340	-	-	5	-	0	0	0	
SZ341	-	-	4	-	0	1xSZ	2	
SZ343	-	-	0	-	0	0	1	
SZ344	-	-	0	-	0	0	1	
SZ346	-	-	3	-	1	0	0	
SZ348	-	-	5	-	0	1xCSZ	1	
SZ349	-	-	2	-	1	1xSZ, 2xCSZ	1	
SZ351	-	-	1	-	0	1xSZ	2	
SZ352	-	-	1	-	0	1xSZ	0	
SZ353	-	-	0	-	1	1xSZ	1	
SZ354	-	-	1	-	0	0	0	
SZ355	15	0	7	1	0	2xCSZ	1	*
SZ356	18	0	5	1	1	1xCSZ	1	*
SZ357	-	-	1	-	0	1xCSZ	0	
SZ358	-	-	3	-	0	0	1	
SZ363	-	-	1	-	2	1xSZ, 1xCSZ	0	
SZ364	14	1	6	4	1	0	0	*
SZ365	-	-	3	-	1	1xSZ, 1xSSZ	3	
SZ366	-	-	5	-	0	1xCSZ	1	
SZ367	-	-	2	-	2	1xSZ	1	
SZ368	-	-	0	-	0	0	1	
SZ369	-	-	1	-	0	0	0	
SZ370	-	-	2	-	0	1xSZ	1	
SZ371	10	0	1	1	0	1xCSZ	1	*
SZ373	-	-	2	-	1	1xCSZ	0	
SZ374	-	-	3	-	0	0	1	
SZ375	-	-	1	-	1	1xSZ	0	
SZ376	-	-	1	-	0	0	0	
SZ377	-	-	3	-	2	2xCSZ	0	
SZ378	-	-	2	-	0	0	1	
SZ379	-	-	1	-	0	1xCSZ	1	
SZ380	-	-	2	-	0	0	4	
SZ381	-	-	2	-	0	0	2	
SZ383	9	0	5	0	0	2xCSZ	2	*

ID	CADD	LoF	Cpht	DNM	EG	CV	LDel	Trio
SZ384	-	-	2	-	1	0	3	
SZ385	-	-	0	-	0	1xC SZ	0	
SZ386	-	-	3	-	0	2xSZ, 2xC SZ	2	
SZ387	20	0	6	2	0	0	1	*
SZ388	-	-	0	-	0	1xC SZ	0	
SZ389	-	-	1	-	0	1xSZ	0	
SZ390	-	-	3	-	0	0	2	
SZ392	-	-	1	-	0	1xSZ, 1xC SZ	0	
SZ393	-	-	2	-	0	2xC SZ	1	
SZ394	-	-	2	-	1	1xC SZ	0	
SZ395	17	0	7	2	0	2xC SZ	4	*
SZ398	14	0	3	0	0	1xC SZ	1	*
SZ399	-	-	3	-	2	1xC SZ	3	
SZ400	-	-	1	-	1	2xC SZ	3	
SZ401	-	-	4	-	0	0	1	
SZ402	-	-	3	-	0	2xC SZ, 1xC SZ	0	
SZ403	-	-	2	-	0	0	2	
SZ404	-	-	3	-	0	0	2	
SZ405	17	1	1	1	0	1xC SZ	1	*
SZ406	-	-	3	-	1	0	1	
SZ407	-	-	3	-	1	0	2	
SZ408	-	-	1	-	0	1xC SZ	3	
SZ411	-	-	1	-	0	1xC SZ	0	
SZ412	-	-	0	-	1	0	0	
SZ413	-	-	1	-	1	0	0	
SZ415	-	-	1	-	1	2xC SZ	1	
SZ416	13	1	1	1	1	0	1	*
SZ418	-	-	2	-	1	0	3	
SZ419	-	-	0	-	0	1xC SZ	2	
SZ420	-	-	1	-	0	0	1	
SZ421	-	-	2	-	0	1xC SZ	0	
SZ422	-	-	3	-	0	0	2	

Table 8.1: Candidate/sample overview. For each affected patient, the number of detected variants/genes is specified. CADD: Genes with increased genetic burden; LoF: Genes with *de novo* knock out; Cpht: Rare damaging compound heterozygous variants; DNM: Rare damaging *de novo* variants; EG: Genes with disrupted protein-interaction sites; CV: Variants with known schizophrenia associations based on ClinVar, SZ: schizophrenia, CSZ: Childhood-onset schizophrenia, SSZ: Susceptibility to schizophrenia; LDel: Amount of long deletion on chromosome 22; Trio(*) indicates whether parents of the patient were sequenced; Finding summarises the most important results for the respective samples.

FURTHER PROJECTS AND PUBLICATIONS

Beside my thesis, I worked on several other side-projects that included various **NGS**-related topics. Next the development of two software frameworks for DNA and RNA analysis, I was also involved in data analysis of two projects focusing on the effect of Epstein-Barr viral miRNAs on $CD4^+$ and $CD8^+$ T cells. In the following sections, I will provide an overview about these projects, my contribution to it, as well as the respective peer-reviewed publications.

9.1 KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis

Experimental labs generate molecular biology data at an ever increasing throughput. Large-scale data exploration is an essential need whenever experimental data need to be compared to public data resources such as human variants. To cope with that, successful Big Data programs need to efficiently use available manpower for conducting the experiments and the analysis, thus providing an infrastructure useful to biologists and data analysts is a critical step. A framework that focuses on this problem is the scientific workflow system [KNIME](#).

In this project, we designed and implemented software solutions within the [KNIME](#) framework to enable easy, standardized and reproducible workflows for [NGS](#) analyses. Together with several colleagues, I created a toolbox with more than 40 specific [KNIME](#) nodes that can execute various processes like simple file operations as well as the most important [KNIME](#) processing steps. Taken together, these modules can be combined into workflows for exome and whole-genome variant calling, splice-site detection as well as differential expression analysis.

Due to the modularity of these [KNIME](#) nodes as well as the underlying Java interface, the package is easily extendable.

Since big data analyses come with specific requirements, we further developed a dedicated high-throughput execution method that is designed for handling [NGS](#) data in large scale. This improvement of the default execution method allows to automatically handle typical errors that arise during the processing of hundreds of unique data sets without time-costly manual intervention.

Overall, the [KNIME4NGS](#) software toolbox can substantially lower the effort for scientists entering into the areas of [NGS](#) and Big Data and provides a basis for a wide range of customized [NGS](#) analysis workflows.



Genome analysis

KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis

Maximilian Hastreiter, Tim Jeske, Jonathan Hoser, Michael Kluge, Kaarin Ahomaa, Marie-Sophie Friedl, Sebastian J. Kopetzky, Jan-Dominik Quell, H.-Werner Mewes and Robert Küffner*

Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on September 7, 2016; revised on December 14, 2016; editorial decision on January 2, 2017; accepted on January 4, 2017

Abstract

Summary: Analysis of Next Generation Sequencing (NGS) data requires the processing of large datasets by chaining various tools with complex input and output formats. In order to automate data analysis, we propose to standardize NGS tasks into modular workflows. This simplifies reliable handling and processing of NGS data, and corresponding solutions become substantially more reproducible and easier to maintain. Here, we present a documented, linux-based, toolbox of 42 processing modules that are combined to construct workflows facilitating a variety of tasks such as DNaseq and RNAseq analysis. We also describe important technical extensions. The high throughput executor (HTE) helps to increase the reliability and to reduce manual interventions when processing complex datasets. We also provide a dedicated binary manager that assists users in obtaining the modules' executables and keeping them up to date. As basis for this actively developed toolbox we use the workflow management software KNIME.

Availability and Implementation: See <http://ibisngs.github.io/knime4ngs> for nodes and user manual (GPLv3 license)

Contact: robert.kueffner@helmholtz-muenchen.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Success of large-scale data analysis depends on sophisticated bioinformatic support to process, integrate, analyze and interpret Big Data volumes. In order to cope with the increased throughput of massive data generating experiments we create structured and reusable processing workflows for various analyses that are easy to apply and can be shared among the community. Although other comparable workflow management tools such as Galaxy (Goecks *et al.*, 2010) exist, we selected the open source platform KNIME (Konstanz information miner, Berthold *et al.* (2008)), as it offers an intuitive graphical user interface, is user friendly and easily extendable. With a very strong and active community, a remarkable number of new functions have been incorporated into KNIME. For instance, Knime4Bio (Lindenbaum *et al.*, 2011) has been developed for the interpretation of biological NGS datasets starting from variant calls.

Here, we describe an extension of KNIME by adding the functionality for essential NGS data processing. We developed a comprehensive linux-based KNIME toolkit including well-documented modules (nodes) for important steps like read pre-processing, read mapping, variant calling, detection of differential expression and annotation. Complementary to previously existing nodes, our toolbox now facilitates the assembly of basic building blocks into a wide range of customized NGS analysis workflows.

2 Implementation

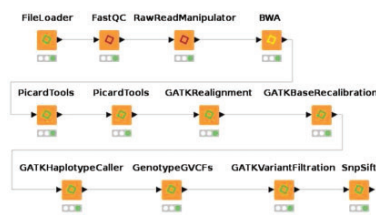
The nodes provided by our extension can be connected smoothly by using their generic interfaces. They are well documented and assist users in the robust configuration of the underlying

Table 1. Collection of tools that are included in the KNIME4NGS extension

<i>Whole exome/genome sequencing</i>	<i>Description</i>
FastQC ^a and RawReadManipulator	Analysis and filtering of raw NGS reads
BWA, Bowtie2 and Segemehl	Mapping against reference genome
Samtools and Picard Utilities	File conversion, PCR duplicate removal as well as several auxiliary functionalities
GATK Utilities	GATK Best Practices and walkers for VCF manipulation
KGGSeq, SNPsift/eff, VEP	Variant annotation and filtration
<i>RNA-Seq</i>	<i>Description</i>
STAR	Splice-aware read aligner
FeatureCount	Assigning sequence reads to features
DESeq, EdgeR and Limma	Detection of differential expressed genes

For complete list and RawReadManipulator see user manual.

^aModified Version.

**Fig. 1.** Exemplary workflow for analysing whole-exome data. This includes all steps from raw read quality control up to variant filtration and annotation

executables. Beyond the individual nodes we also added new layers of functionality that support the management and update of the underlying software packages as well as their high-throughput execution.

2.1 KNIME nodes

Our KNIME node extension for NGS currently contains 42 nodes (excerpt shown in Table 1. See comprehensive user manual for complete list and full description). Besides a number of auxiliary nodes, this extension provides building blocks and wrappers for well-established tools like BWA (Li and Durbin, 2010), DESeq (Anders and Huber, 2010) and GATK (McKenna et al., 2010). The nodes can be easily combined to create standardized workflows starting from initial preprocessing of raw data, up to the variant calling and biological annotation (Fig. 1). This enables expert and non-expert users to set up reliable processing workflows quickly without the need of dealing with command-line tools. We also extended KNIME to use NGS specific file types. This improves workflow robustness enabling automated checks of tool configuration parameters and versions. To organize the required software binaries, we developed a lightweight binary management system accessible via the internal preference page. Thus, distributed nodes do not need to include the underlying executables, which simplifies keeping track with frequently updated software packages. The binary manager (See manual section 6) obtains necessary executables and integrates them into the workflow, minimizing repetitive node configuration.

NGS datasets are typically very large. Parallelizing the workflows and controlling data and results is crucial. Therefore we designed and tested our toolkit to work with corresponding KNIME extensions like openBIS (Bauch et al., 2011), for storing and managing datasets, or the KNIME parallel chunk environment and cluster execution, for high-performance computing.

2.2 High-throughput executor

In the analyses of large datasets, even established tools are prone to spurious premature termination. This aborts the entire workflow and requires extensive human intervention to ensure that all of the parallel executing nodes/samples finish successfully. To reduce this effort as much as possible, we developed the High-Throughput Executor (HTE) as extension of the standard KNIME NodeModel. The HTE model collects process termination data of our nodes (e.g. error logs, execution time) and stores them in a local database provided by the system. Depending on the configuration and the process termination status, it automatically retries the execution of the failed node. The database allows the user to keep track and reduce the effects of unexpected software behaviour caused by for instance insufficient memory or randomly occurring errors. Additionally, the HTE model ensures by dedicated lock-files that not the entire workflow, but only those nodes are re-executed that depend on failed previous steps.

3 Discussion

Analysis of Big Data is often difficult to maintain and reuse as it depends on the correct application of multiple processing steps. As a solution, compiling the necessary tools into standardized graphical workflows makes the analysis pipeline more explicit, transparent and adaptable and can therefore stimulate collaboration between computational and wet lab biologists. The publication of scientific results together with the generating workflows furthermore has the potential to improve representation, reproducibility and dissemination of findings substantially.

We have presented a modular toolkit for the construction of NGS data processing workflows based on the KNIME environment that, in several ways, extends generic solutions e.g. the SeqAn KNIME library (Döring et al., 2008). First of all, our toolkit comprises a set of nodes for individual NGS processing steps that can be combined into various workflows to serve as an extensible basis for many advanced NGS projects. The provided nodes and derived workflows are suitable for experts as well as less experienced users due to their integrated automatic parameter validation and comprehensive node descriptions accessible through the KNIME user interface. The modularity of our nodes allows fast and easy workflow adjustments by adding nodes or creating alternate paths. NGS projects are now characterized by their huge data volume and their need for massive parallel data processing. To improve the smooth handling of such Big Data workflows, we also provide a set of crucial technical KNIME extensions. These improve the robust setup, maintenance and configuration of workflows as well as their reliable execution and debugging with minimal human intervention. Taken together, KNIME4NGS can substantially lower the effort for scientists entering into the areas of NGS and Big Data.

Conflict of Interest: none declared.

References

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, 11, R106.

9.1. KNIME4NGS: A COMPREHENSIVE TOOLBOX FOR NEXT GENERATION SEQUENCING ANALYSIS

- Bauch, A., *et al.* (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12, 1.
- Berthold, M.R. *et al.* (2008) KNIME: The Konstanz Information Miner. In: Preisach, C. *et al.* (eds), *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. Springer Berlin Heidelberg, pp. 319–326.
- Döring, A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9, 11.
- Goecks, J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11, 1.
- Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26, 589–595.
- Lindenbaum, P., *et al.* (2011) Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinformatics*, 27, 3200–3201.
- McKenna, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20, 1297–1303.

9.2 Effect of Epstein-Barr viral miRNAs on $CD4^+$ and $CD8^+$ T cells

Infection with Epstein-Barr virus (EBV) affects most humans world- wide and persists life-long in the presence of robust virus-specific T-cell responses. In both immunocompromised and some immunocompetent people, EBV causes several cancers and lymphoproliferative diseases. EBV encodes at least 44 microRNAs, most of which are expressed in EBV-transformed B cells. However their functions are largely unknown. Within two related projects, in which I conducted bioinformatic analysis of the available NGS data sets, we investigated the effect of the viral miRNAs on $CD4^+$ and $CD8^+$ T cells.

EBV miRNAs inhibit antiviral $CD4^+$ T cells

EBV miRNAs collectively and specifically suppress release of proinflammatory cytokines such as IL-12, repress differentiation of naive $CD4^+$ T cells to Th1 cells, interfere with peptide processing and presentation on HLA class II, and thus reduce activation of cytotoxic EBV-specific $CD4^+$ effector T cells and killing of infected B cells.

Our findings suggest that by rapidly expressing miRNAs, which are themselves nonimmunogenic, EBV counteracts recognition by $CD4^+$ T cells and establishes a program of reduced immunogenicity in recently infected B cells, allowing the virus to express viral proteins required for establishment of life-long infection.

$CD8^+$ T cells surveillance is reduced by EBV miRNAs

Based on the findings received on $CD4^+$ T cells, we investigated whether EBV miRNAs also counteract surveillance by $CD8^+$ T cells. Here we found that EBV miRNAs strongly inhibit recognition and killing of infected B cells by EBV-specific $CD8^+$ T cells. This is achieved throughout multiple mechanisms. EBV miRNAs directly target the peptide transporter subunit TAP2 and reduce levels of the TAP1 subunit, MHC class I molecules, and EBNA1, a protein expressed in most forms of EBV latency and a target of EBV-specific $CD8^+$ T cells. Moreover, miRNA-mediated down-regulation of the cytokine IL-12 decreases the recognition of infected cells by EBV-specific $CD8^+$ T cells.

These distinct pathways ensure that the viral miRNAs are able to evade surveillance of $CD4^+$ and antiviral $CD8^+$ T cells.

Epstein-Barr viral miRNAs inhibit antiviral CD4⁺ T cell responses targeting IL-12 and peptide processing

Takanobu Tagawa,^{1,2*} Manuel Albanese,^{1,2*} Mickaël Bouvet,^{1,2} Andreas Moosmann,^{1,2} Josef Mautner,^{1,2,3} Vigo Heissmeyer,^{4,5} Christina Zielinski,⁶ Dominik Lutter,⁷ Jonathan Hoser,⁸ Maximilian Hastreiter,⁸ Mitch Hayes,⁹ Bill Sugden,⁹ and Wolfgang Hammerschmidt^{1,2}

¹Research Unit Gene Vectors, Helmholtz Zentrum München, German Research Center for Environmental Health and ²German Centre for Infection Research (DZIF), Partner site Munich, Germany, D-81377 Munich, Germany

³Children's Hospital, Technical University Munich, D-80337 Munich, Germany

⁴Research Unit Molecular Immune Regulation, Institute of Molecular Immunology, Helmholtz Zentrum München, German Research Center for Environmental Health Munich and ⁵Institute for Immunology, University of Munich, D-80539 Munich, Germany

⁶Institute for Medical Microbiology, Immunology and Hygiene, Technical University Munich, D-80337 Munich, Germany

⁷Institute for Diabetes and Obesity, Helmholtz Zentrum München and ⁸Institute of Bioinformatics and System Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, D-85764 Munich, Germany

⁹McArdle Laboratory for Cancer Research, University of Wisconsin-Madison, Madison, WI 53706

Epstein-Barr virus (EBV) is a tumor virus that establishes lifelong infection in most of humanity, despite eliciting strong and stable virus-specific immune responses. EBV encodes at least 44 miRNAs, most of them with unknown function. Here, we show that multiple EBV miRNAs modulate immune recognition of recently infected primary B cells, EBV's natural target cells. EBV miRNAs collectively and specifically suppress release of proinflammatory cytokines such as IL-12, repress differentiation of naive CD4⁺ T cells to Th1 cells, interfere with peptide processing and presentation on HLA class II, and thus reduce activation of cytotoxic EBV-specific CD4⁺ effector T cells and killing of infected B cells. Our findings identify a previously unknown viral strategy of immune evasion. By rapidly expressing multiple miRNAs, which are themselves nonimmunogenic, EBV counteracts recognition by CD4⁺ T cells and establishes a program of reduced immunogenicity in recently infected B cells, allowing the virus to express viral proteins required for establishment of life-long infection.

INTRODUCTION

EBV is both ubiquitous and immunogenic. This oncogenic herpesvirus (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2010) has evolved multiple genes to fend off immune responses when its infection is established (Hislop et al., 2002; Rowe et al., 2007; Rensing et al., 2008; Zuo et al., 2009; Qiu et al., 2011; Rancan et al., 2015). Despite these measures, EBV-specific T cells constitute a considerable fraction of the memory T cell repertoire of the latently infected human host (Hislop et al., 2002) and are essential in controlling latent EBV infection (Moosmann et al., 2010). In fact, immunocompromised patients have an increased incidence of EBV-associated malignancies (Gottschalk et al., 2005).

EBV infects nondividing B lymphocytes, activates them, and drives them to proliferate, thus amplifying the load of viral genomes. Once activated, infected B cells acquire properties of antigen-presenting cells. After infection, they rapidly present epitopes of structural proteins from incoming virus particles and transiently express lytic genes that are otherwise

characteristic of EBV's productive cycle (Kalla and Hammerschmidt, 2012). This prelatent phase of infection includes expression of two genes coding for viral immunoevasins, BNLF2a and BCRF1 (Jochum et al., 2012), which inhibit the recognition of the infected cells by EBV-specific effector T cells and natural killer cells, respectively. These two viral proteins are insufficient, however, to overcome T cell recognition (Jochum et al., 2012). Within 7–10 d, EBV establishes a latent infection in the infected B cells and expresses only few or no viral genes, which reduces their risk of becoming eliminated by the immune-competent host.

Thus, early infection could be EBV's Achilles heel, a window when the infected cell expresses and presents many viral antigens to immune cells but is inadequately protected from the host's immune response. We have now established that EBV's miRNAs overcome this vulnerability; they protect newly infected B lymphocytes from immune eradication by CD4⁺ T cells, supporting EBV's lifelong success.

EBV encodes at least 44 microRNAs (miRNAs; Barth et al., 2011), which are small RNA regulatory molecules of ~22 nt in length (Bartel, 2004). miRNAs encoded by herpesviruses

*T. Tagawa and M. Albanese contributed equally to this paper.

Correspondence to Wolfgang Hammerschmidt: hammerschmidt@helmholtz-muenchen.de

Abbreviations used: CTSB, cathepsin B; LCL, lymphoblastoid cell line; LMP, latent membrane protein; miRNA, microRNA; RISC, RNA-induced silencing complex.

© 2016 Tagawa et al. This article is distributed under the terms of an Attribution-NonCommercial-Share Alike-No Mirror Sites license for the first six months after the publication date (see <http://www.rupress.org/terms>). After six months it is available under a Creative Commons License (Attribution-NonCommercial-Share Alike 3.0 Unported license, as described at <http://creativecommons.org/licenses/by-nc-sa/3.0/>).



JEM

are reported to play important roles in cell proliferation, development, immune regulation, and apoptosis in infected cells (Skalsky and Cullen, 2010). The EBV-encoded miRNAs have been found to control expression of several cellular genes with antiapoptotic functions, but they also reportedly down-regulate *MICB* (Nachmani et al., 2009), *CXCL11* (Xia et al., 2008), and *NLRP3* (Haneklaus et al., 2012) and thus interfere with innate immune responses and inflammation. Interestingly, *MICB*, a gene encoding a ligand for the activating receptor NKG2D expressed on T and NK cells, is also targeted by miRNAs of Kaposi sarcoma-associated herpesvirus and human cytomegalovirus (Nachmani et al., 2009; Grundhoff and Sullivan, 2011). These studies imply that certain miRNAs encoded by herpesviruses target pathways involved in innate immune recognition.

EBV's miRNAs have been studied by several groups with established, EBV-infected cell lines obtained from biopsies of nasopharyngeal carcinoma and Burkitt's lymphoma, or lymphoblastoid cell lines (LCLs) derived from infecting primary B lymphocytes with EBV in vitro (Dölken et al., 2010; Gottwein et al., 2011; Kuzembayeva et al., 2012; Riley et al., 2012; Erhard et al., 2013). High-throughput target screens using immunoprecipitation of the RNA-induced silencing complex (RISC) and deep sequencing have identified many potential targets of EBV miRNAs, but the catalogs of predicted targets assembled by different groups have a surprisingly small overlap (Klinke et al., 2014). This lack of consensus may be due to the accumulation of profound differences in gene expression between different long-term, cultivated EBV-infected cell lines that do not reflect the impact of EBV's miRNAs in vivo.

To circumvent these problems, we developed an experimental approach using primary human B lymphocytes, and analyzed them during their initial days of EBV infection (Seto et al., 2010; Vereide et al., 2014). We infected the B lymphocytes with two EBV strains with and without miRNA genes, compared the gene expression in the infected cells, and examined them for their immune recognition.

We found that EBV-encoded miRNAs regulated several immune pathways, which affected CD4⁺ T cell differentiation and activation. In addition, key molecules important for interactions with CD4⁺ T cells were down-regulated. EBV miRNAs repressed the secretion of IL-12, which resulted in suppression of type 1 helper T cell (Th1) differentiation. Viral miRNAs controlled gene expression of HLA class II and three lysosomal enzymes important for proteolysis and epitope presentation to CD4⁺ T cells. Such a wholesale inhibition of adaptive immune responses by multiple miRNAs of a single pathogen is unprecedented. Our findings explain the abundance of miRNAs in complex persisting viruses, and clarify how EBV can escape elimination for the lifetime of its host in spite of intense adaptive immune responses.

RESULTS

EBV miRNAs control immune regulatory pathways

We searched for cellular targets of EBV's miRNAs, using an experimental system that closely mimics human infection

in vivo. Two strains of EBV, a laboratory strain (wt/B95-8) that expresses 13 miRNAs, and its derivative (Δ miR) that expresses none (Seto et al., 2010) were used to infect freshly isolated B lymphocytes from six donors. We used carefully titrated virus stocks and infected the cells with optimal doses of both viruses (Steinbrück et al., 2015). No differences in the percentage of infected cells were seen when comparing cells infected with wt/B95-8 versus Δ miR EBV. RNAs were isolated on day 5 after infection and sequenced (available from GEO under accession no. GSE75776; see Materials and methods). Genes that were differentially expressed in cells infected with wt/B95-8 versus Δ miR EBV were identified with those having an absolute z -score >1.6 (Fig. 1 A and Table S1). These genes included the published miRNA targets *LY75/DEC205* (Skalsky et al., 2012) and *IPO7* (Dölken et al., 2010). Genes that were consistently down-regulated in wt/B95-8 EBV-infected cells were grouped according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway categories (Fig. 1 B). Down-regulated genes were predominant in pathways linked to apoptosis, cell cycle regulation, and p53 signaling, which were previously proposed to be regulated by EBV miRNAs (Seto et al., 2010; Feederle et al., 2011a,b; Vereide et al., 2014). Unexpectedly, EBV's miRNAs also regulated a wide array of genes with functions in immunity, such as cytokine-cytokine receptor interactions, antigen processing, and HLAs and co-stimulatory molecules (Fig. 1, B and C; and Table S1). We immunoprecipitated RISC (RISC-IP) and found that 14.5% ($\pm 2.4\%$ SD) of all miRNAs were of viral origin in wt/B95-8 EBV-infected cells, dominated by miRNAs of the BHRF1 gene cluster (Fig. 1 D). No appreciable viral miRNA reads were found in cells infected with Δ miR EBV (Fig. 1 D), suggesting that the B lymphocytes of six donors were free of EBV field strains. In wt/B95-8 EBV-infected cells, we detected viral miRNAs as early as day 1 after infection, which reached high levels 5 days post infection (dpi; Fig. 1 E). In RISC-IP, detection of miRNAs was variable among infected B cells of the different donors, a phenomenon that was reported earlier using a related model of established infection and PAR-CLIP experiments (Skalsky et al., 2012; GEO accession no. GSE41437). Therefore, we focused our analyses on candidate mRNAs that were uniformly regulated in all samples (Fig. 1 C), and used RISC-IP results to confirm them (Table S1).

EBV miRNAs inhibit secretion of proinflammatory cytokines and antigen presentation

We confirmed that EBV's miRNAs regulate cytokines central to immune function. Supernatants from B cells infected with the two strains of EBV were assayed for the levels of IL-6, IL-10, TNF, IL12B (IL-12p40), IL-12 (p35/p40), and IL-23 (p19/p40). We added CpG DNA, which stimulates TLR9, for the detection of IL-6 secreted from EBV-infected cells (Iskra et al., 2010). B cells infected with wt/B95-8 EBV secreted less IL-6, TNF, and IL-12p40 than B cells infected with Δ miR EBV. In contrast, release of the anti-inflamma-

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

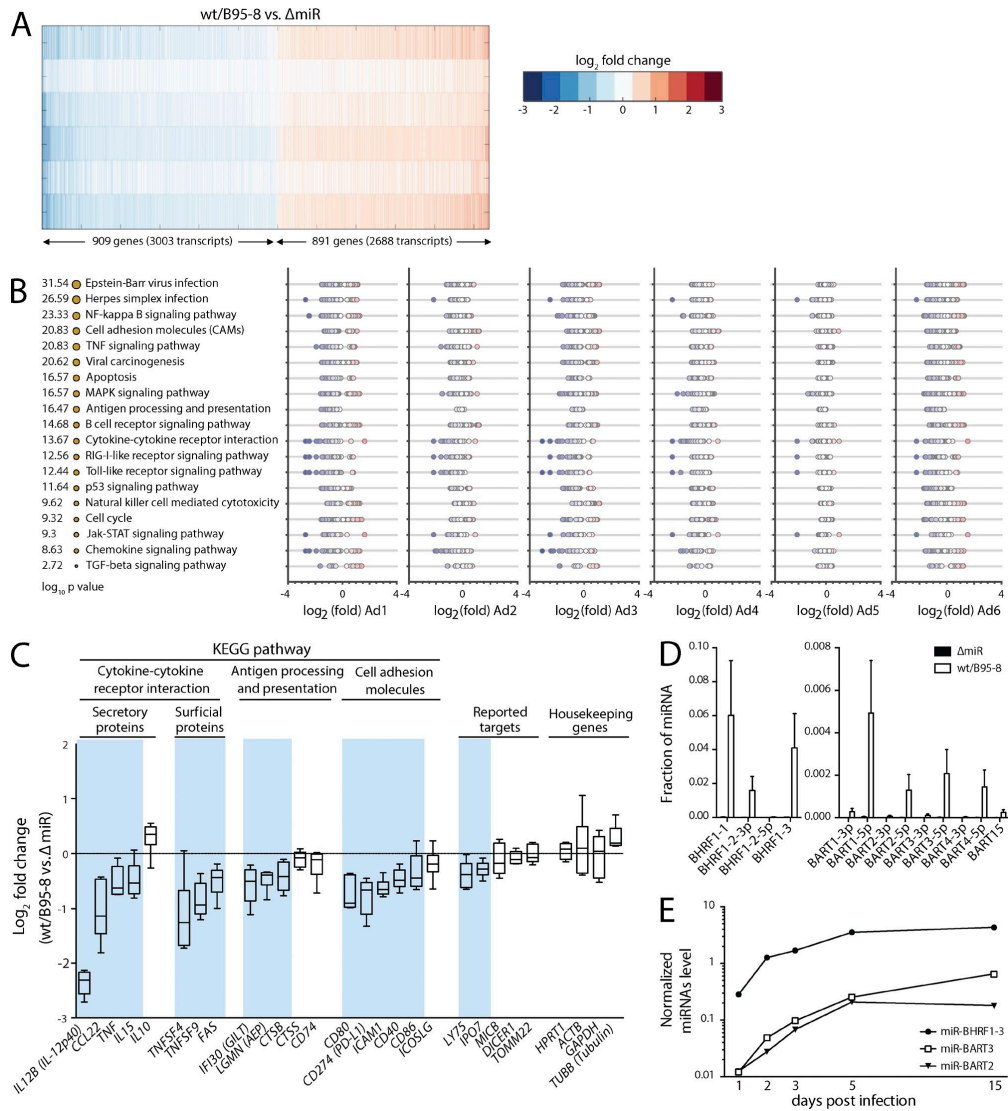


Figure 1. EBV miRNAs affect major pathways of immunity. (A) A heat map of the most strongly regulated genes in wt/B95-8 or Δ miR EBV-infected B cells of six donors (donor Ad1-Ad6) 5 dpi shows differentially expressed gene transcripts with absolute z-scores >1.6. Blue and red indicate down- and up-regulated transcripts, respectively, in wt/B95-8 compared with Δ miR EBV-infected cells. (B) Shown are gene functions according to KEGG pathway categories with the identified pathways sorted by statistical significance. The sizes of the orange dots indicate $-\log_{10}$ P value scores. For each of the six donors, fold change values of differentially expressed transcripts are plotted. As in A, blue and red indicate down- or up-regulation by EBV miRNAs, respectively. Enrichment of specific pathways was estimated by a hypergeometric distribution test via the KEGG API Web service. (C) Inhibition of selected transcripts associated with adaptive immune responses is shown together with previously reported targets of EBV miRNAs and common housekeeping genes. Blue background shadings indicate genes down-regulated by viral miRNAs (Table S1). (D) Levels of indicated EBV miRNAs in B cells infected with wt/B95-8 or Δ miR EBV-infected B cells of six donors (donor Ad1-Ad6) 5 dpi were quantified by RISC-IP-seq. Mean \pm SD are shown. (E) Three miRNAs, which represent different primary miRNA transcripts in EBV-infected B cells, were quantified with stem-loop qPCR over time. One of two independent experiments is shown.

JEM

tory cytokine IL-10 appeared to be unaffected by viral miRNAs (Fig. 2 A) consistent with our transcriptome analysis (Fig. 1 C). Secretion of IL-12 (p35/p40 or IL-12p70) and IL-23 (p19/p40), both of which contain the IL-12p40 subunit (Szabo et al., 2003), encoded by the *IL12B* gene, was significantly reduced in wt/B95-8 EBV-infected cells compared with Δ miR EBV-infected cells (Fig. 2 A). Viral miRNAs also inhibited the secretion of IL-12p40 from PBMCs infected with wt/B95-8 EBV (Fig. 2 B). IL-12p40 secretion from PBMCs infected with Δ miR EBV was reduced when B cells were removed from the PBMCs, indicating that B cells are the main contributors to release of IL-12p40 in PBMCs. Remarkably, our transcriptome analysis revealed the consistent reduction of *IL12B* mRNA with EBV's miRNAs, reducing it by 80% in all six donors' B lymphocytes (Fig. 2 C). Quantitative RT-PCR confirmed this finding (Fig. 2 D).

Multiple EBV miRNAs target *IL12B* and prevent Th1 differentiation of naive CD4⁺ T cells

We investigated whether *IL12B* was a direct target of EBV miRNAs. EBV's miR-BART1, miR-BART2, and miR-BHRF1-2 repressed the luciferase activity of the *IL12B* reporter (Fig. 3 A). The mutation of predicted binding sites of miR-BART1, miR-BART2, or miR-BHRF1-2 abrogated their ability to inhibit the *IL12B* reporter (Fig. 3 A and Fig. S1), confirming the direct control of *IL12B* by these miRNAs. We similarly analyzed miR-BART10 and miR-BART22, which are present in field strains of EBV but not in wt/B95-8 EBV. For these miRNAs, mutations of their predicted target sites only partially relieved inhibition (Fig. 3 A and Fig. S1), suggesting the presence of additional binding sites in the *IL12B* transcript. In summary, these experiments validated *IL12B* as a direct target of multiple viral miRNAs.

IL-12 is critical for differentiation of Th1 cells (Szabo et al., 2003). Therefore, we co-cultured naive CD4⁺ T cells with autologous EBV-infected B cells (Fig. 3 B). Relative to Δ miR EBV, wt/B95-8 EBV-infected B cells repressed Th1 differentiation (Fig. 3, C and D). An antibody that neutralizes the functions of IL12B, but not an isotype control antibody, suppressed Th1 differentiation when T cells were co-cultured with Δ miR EBV-infected cells (Fig. 3 E), indicating that IL-12, secreted from EBV-infected and activated B cells, was responsible for generation of Th1 cells. Thus, EBV miRNAs suppress the release of IL-12 from infected cells and thereby interfere with formation of Th1 cells, which are important antiviral effectors.

Viral miRNAs directly and indirectly control antigen presentation

Having shown that EBV miRNAs interfere with CD4⁺ T cell differentiation, we turned our attention to molecules that are involved in recognition of infected cells by specific CD4⁺ T cells. We quantified levels of surface proteins with a role in HLA class II antigen presentation (Fig. 4). All three subclasses of HLA class II (HLA-DR, HLA-DQ, and HLA-DP) tested were reduced in wt/B95-8 relative to Δ miR EBV-

infected B cells (Fig. 4, A and B), as were many co-stimulatory and adhesion molecules 5 and 15 dpi (Fig. 4 C). MHC class II molecules were also affected, but to a lesser extent.

Among the many co-receptors and adhesion molecules down-regulated in cells infected with wt/B95-8 EBV (Fig. 4 C), we searched for direct targets of EBV miRNAs but found only *CD40* and *FAS* in RISC-IPs (Table S1), which we could not confirm in subsequent luciferase assays using miRNAs encoded by wt/B95-8 EBV. Interestingly, the viral latent membrane protein 1 (LMP1) activates the CD40 pathway, inducing important immune co-receptors (Kieser and Sterz, 2015), but several viral BART miRNAs were reported to control LMP1 expression (Lo et al., 2007; Riley et al., 2012; Verhoeven et al., 2016). We tested these findings in our model of newly infected B cells, and found reduced but highly variable levels of LMP1 transcripts (Fig. 4 D) and protein (Fig. 4 E) in B cells infected with wt/B95-8 EBV compared with Δ miR EBV 5 dpi. We identified miR-BART3 and miR-BART16 as inhibiting LMP1 in reporter assays (Fig. 4 F). miR-BART3 is encoded in wt/B95-8 EBV, whereas miR-BART16 (Fig. 4 F) is only present in field strains of EBV. We also tested miR-BART1 and miR-BART17, which were reported together with miR-BART16 to target LMP1 3'-UTR (Lo et al., 2007), but failed to confirm that miR-BART1 and miR-BART17 target LMP1 (Fig. 4 F). Collectively, our results showed that viral miRNAs limit LMP1 gene expression and thereby indirectly inhibit surface expression of some immune co-receptors and adhesion molecules.

Viral miRNAs target lysosomal enzymes and inhibit antigen processing

According to our transcriptome analysis, genes encoding lysosomal enzymes actively involved in MHC class II peptide processing (Blum et al., 2013) were inhibited by EBV miRNAs (Fig. 1 C). These included *IFI30* (coding for IFN- γ -regulated thiol reductase GILT), *LGMN* (coding for asparagine endopeptidase AEP alias legumain), and *CTSB* (coding for the peptidase cathepsin B). Expression of all three genes was reduced by EBV miRNAs (Fig. 1 C), which we verified by quantitative RT-PCR (Fig. 5 A). We found that EBV's miR-BART1, miR-BART2, and miR-BHRF1-2 could directly regulate *IFI30*, *LGMN*, and *CTSB* gene expression via their 3'-UTRs in luciferase reporter assays (Fig. 5 B and Fig. S1). Importantly, the knock-down of these three genes (Fig. 5 C) resulted in reduced antigen presentation of exogenously loaded protein (Fig. 5 D). Collectively, our results show that EBV miRNAs interfere with processes involved in MHC class II antigen presentation at multiple levels, including lysosomal protein degradation, HLA class II expression, and co-stimulatory molecule expression.

EBV miRNAs inhibit recognition of infected B cells by EBV-specific CD4⁺ T cells

Next, we asked whether these multiple levels of regulation ultimately resulted in reduced MHC class II-mediated recognition

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

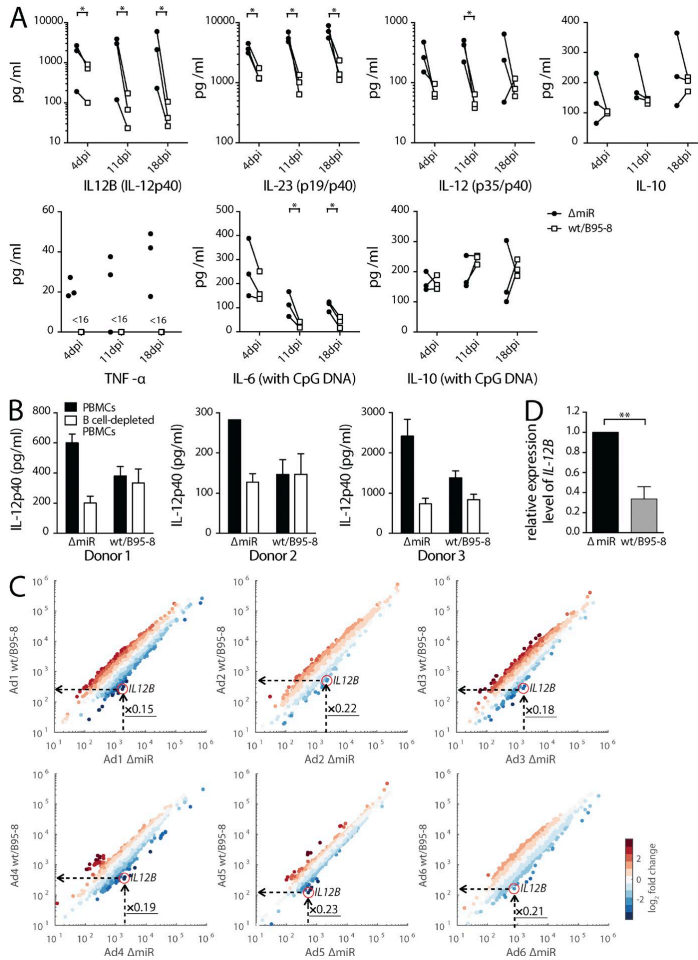


Figure 2. EBV miRNAs inhibit secretion of proinflammatory cytokines. (A) B cells infected with wt/B95-8 or Δ miR EBV for 4, 11, or 18 d were cultivated for an additional 4 d to determine levels of selected cytokines by ELISA ($n = 3$). CpG DNA was added where indicated. Paired samples from individual donors are connected by solid lines. P values were calculated by a paired two-tailed t test. <16 , under the detection limit (16 pg/ml); *, $P < 0.05$. (B) PBMCs or PBMCs depleted of B cells ($n = 3$) were infected with either wt/B95-8 or Δ miR EBV, and concentrations of IL-12p40 in the supernatants of the infected B cells after 5 d were determined by ELISA. (C) Shown are scatter plots of transcriptomes of B cells from six donors infected with wt/B95-8 or Δ miR EBV for 5 d. Fold changes of transcript levels are indicated as blue or red dots, indicating down- and up-regulated transcripts, respectively. The individual *IL12B* transcripts are highlighted by red circles and the calculated fold changes (x -values) are provided. (D) Transcript levels of *IL12B* were measured with quantitative RT-PCR in RNA preparations of B cells infected with wt/B95-8 or Δ miR EBV for 5 d ($n = 4$). **, $P < 0.01$.

of EBV-infected cells by antiviral CD4⁺ T cells. CD4⁺ T cells from EBV-positive individuals were enriched for EBV-specific T cells by repeated stimulation with irradiated wt/B95-8 EBV-infected autologous LCLs. The EBV-specific CD4⁺ T cells were then co-cultured with autologous B cells that had been infected with the two EBV strains 5 d earlier (Fig. 6 A, top). Release of IFN- γ by EBV-specific CD4⁺ T cells was substantial when co-cultured with Δ miR EBV-infected cells as targets, but was consistently reduced when co-cultured with wt/B95-8 EBV-infected B cells at all cell ratios tested (Fig. 6 B). Activation of EBV-specific CD4⁺ T cells, measured as IFN- γ release, was observed in autologous and partially matched but not in HLA-mismatched conditions (Fig. 6 C and Table S2), indicating that the activation was HLA class II-restricted.

We also tested an antigen-specific CD4⁺ T cell clone (Fig. 6, A [bottom] and D) directed against the FGQ peptide, an epitope derived from the viral glycoprotein gp350 (Adhikary et al., 2006). We observed dramatically reduced T cell activities with target B cells infected with wt/B95-8 EBV compared with Δ miR EBVs 5 dpi (Fig. 6 D). T cell activities were much reduced at 15 dpi, but a difference between wt/B95-8 and Δ miR EBV-infected target cells remained detectable. Weak recognition on day 15 is in accordance with gp350 protein being delivered as a component of the virion (Adhikary et al., 2006) but not synthesized during prelatency or latency (Kalla et al., 2010). Interestingly, expression of cell surface HLA class II levels peaked between 4 to 10 dpi (Fig. 6 E) suggesting the importance of

JEM

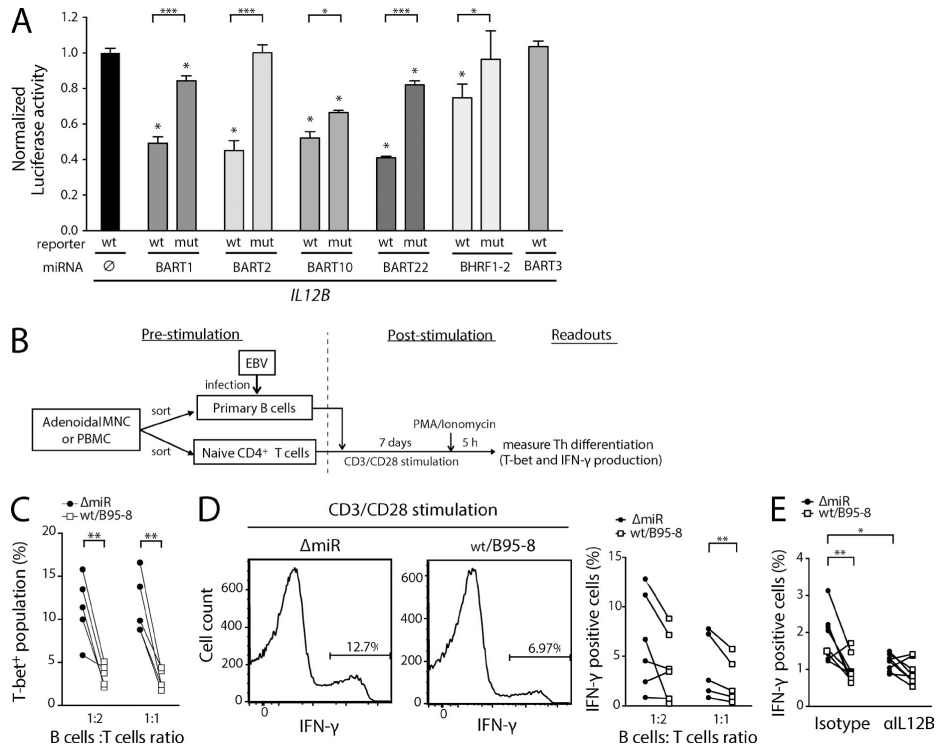


Figure 3. EBV miRNAs inhibit *IL12B* directly and prevent Th1 differentiation. (A) HEK293T cells were cotransfected with miRNA expression vectors and luciferase reporter plasmids carrying a wild-type or mutated 3'-UTR (Fig. S2) as indicated ($n = 3$). The luciferase activities were normalized to lysates from cells cotransfected with the wild-type 3'-UTR reporter and an empty plasmid. wt, wild-type 3'-UTR; mut, mutated 3'-UTR; ∅, empty plasmid. P-values were calculated by an unpaired two-tailed Student's *t* test. *, $P < 0.05$; ***, $P < 0.001$, with respect to the luciferase activity of the wild-type reporter cotransfected with empty plasmid. (B) Schematic representation of the steps for experiments shown in C and D. Primary B cells sorted from adenoids or PBMCs were infected with either wt/B95-8 or EBV Δ miR EBV and co-cultured with autologous naive CD4⁺ T cells, which were stimulated with α CD3/ α CD28 antibody-conjugated beads for 7 d. Th1 differentiation was assessed by intracellular staining of T-bet and IFN- γ after stimulation with PMA/ionomycin for 5 h. (C and D) Naive CD4⁺ T cells were cultivated for 7 d with autologous, newly infected B cells and α CD3/ α CD28 antibody-conjugated beads at indicated ratios ($n = 5-6$). Proliferating PMA- and ionomycin-restimulated Th1 cells were quantified by intracellular T-bet (C) and IFN- γ (D) staining. (D, left) Representative flow cytometry analyses; (right) summary of all experiments. Solid lines indicate paired samples from five to six individual donors. (E) Naive CD4⁺ T cells were cocultivated with wt/B95-8 or Δ miR-infected B cells at a B/T cell ratio of 1:1 ($n = 8$) as shown in (B). An anti-IL12B antibody was administered at a concentration of 5 μ g/ml, and an irrelevant antibody of the same isotype was used as a control. Solid lines indicate paired samples from individual donors. *, $P < 0.05$; **, $P < 0.01$.

viral miRNAs that counteract CD4⁺ T cell recognition in the early days of infection.

EBV-specific CD4⁺ T cells have cytolytic activity (Adhikary et al., 2006). In allogeneic, partially HLA-matched conditions, EBV-specific CD4⁺ T cells consistently showed stronger cytotoxicity of target B cells infected with Δ miR EBV than cells infected wt/B95-8 EBV (Fig. 6 F).

Collectively, we have discovered that EBV miRNAs inhibit the recognition and elimination of infected B cells by HLA class II-restricted CD4⁺ T cells. Apparently, EBV uti-

lizes multiple miRNAs to interfere with proinflammatory cytokines, antigen processing, and epitope presentation of the infected B lymphocyte to evade EBV-specific and antiviral CD4⁺ T cell responses early after infection.

DISCUSSION

EBV infects its human hosts for their lifetime, residing in nonproliferating B cells largely invisible to the host's immune response (Thorley-Lawson, 2005). EBV, to be the successful pathogen that it is, however, must both establish a latent in-

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

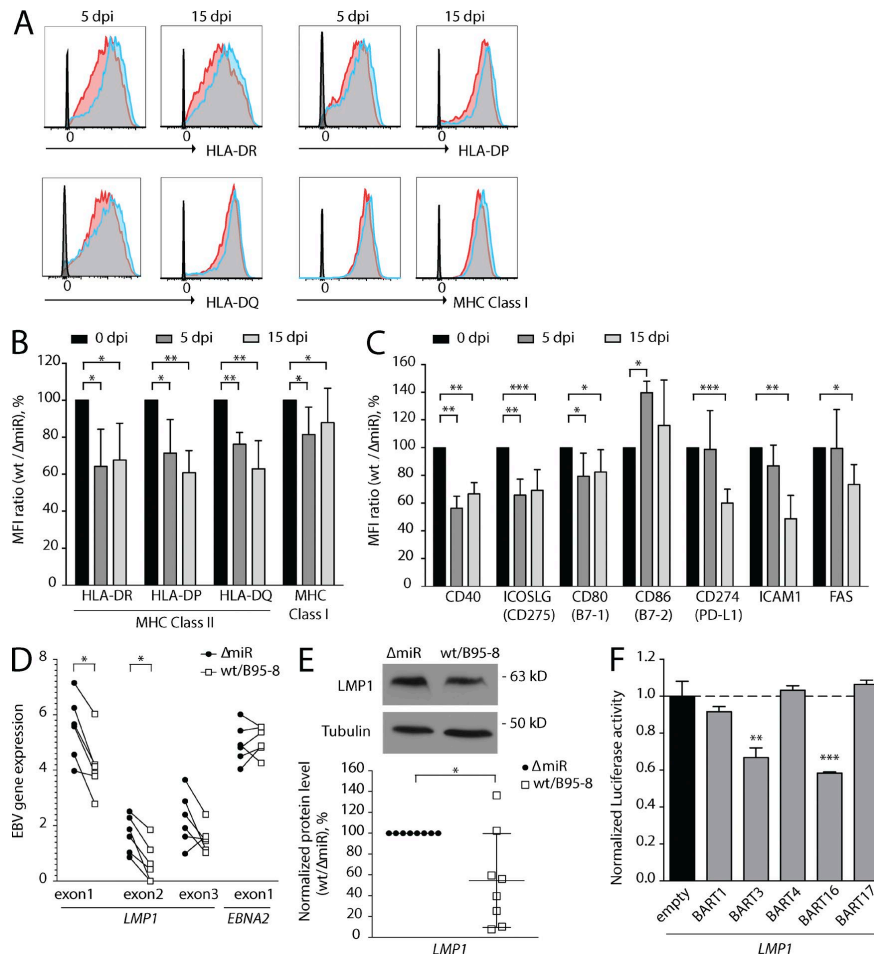


Figure 4. EBV miRNAs control cell surface levels of HLAs and co-receptors. (A) FACS panels show the expression profiles of three HLA class II gene families and HLA class I protein on B cells on 5 dpi. One representative example is shown as a histogram for each condition. (B and C) Cell surface expression of HLA molecules (B) and of co-stimulatory and adhesion molecules (C) was measured after immunostaining for proteins inhibited by EBV miRNAs. Ratios (wt/B95-8 divided by ΔmiR EBV-infected B cells) are shown as median fluorescence intensity (MFI). Means ± SD of experiments with infected B cells from 5–10 donors are shown. P-values were calculated by a paired two-tailed Student's *t* test. *, *P* < 0.05; **, *P* < 0.01; ***, *P* < 0.001. (D) Viral gene expression obtained from the transcriptome data shown in Fig. 1 C was quantified for exons to analyze splicing variants precisely. The y-axis shows natural-log values (*n* = 6) with paired samples from individual donors connected by solid lines. Statistical significance was assessed with the repeated-measurements ANOVA. *, adjusted *P* < 0.01. (E) Cell lysates were prepared from B cells infected with wt/B95-8 or ΔmiR EBV for 5 d and analyzed by Western blotting for expression of LMP1 and tubulin. An example (top) and the quantification of all results (bottom) are shown. Protein levels were measured relative to tubulin and LMP1 levels in ΔmiR EBV-infected cells were set to 100%. Mean ± SD are shown (*n* = 8). (F) Dual luciferase reporter assays with LMP1 3'-UTR are shown (*n* = 3). P-values were calculated by an unpaired two-tailed Student's *t* test. *, *P* < 0.05; **, *P* < 0.01; ***, *P* < 0.001.

fection and produce and disseminate progeny virus, all in the face of robust innate and adaptive immune responses.

Such responses include EBV-specific CD4⁺ T cells, which have an important role in controlling EBV infection

and disease. For example, patients with EBV-associated tumors treated with virus-specific T cell preparations showed better clinical responses if the preparations contained larger fractions of CD4⁺ T cells (Haq et al., 2007; Icheva et al.,

JEM

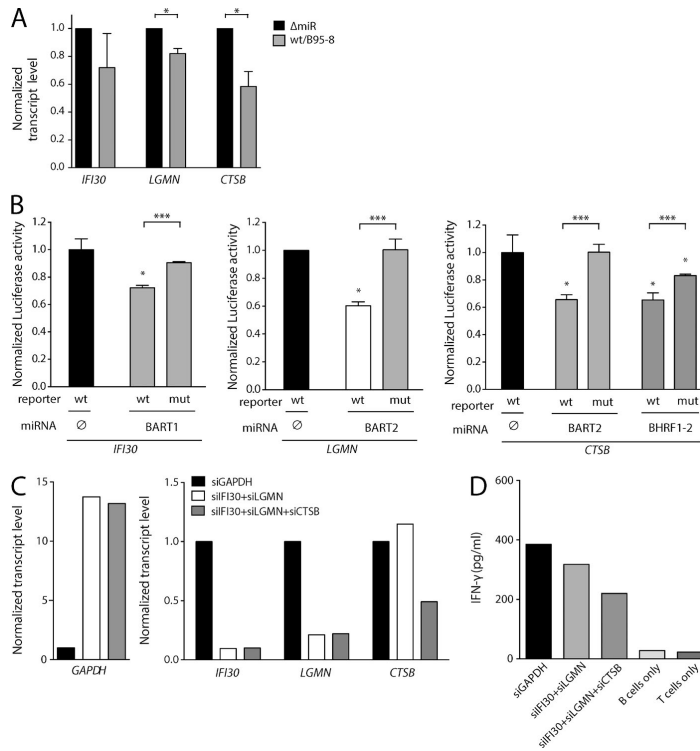


Figure 5. Viral miRNAs inhibit two lysosomal endopeptidases and a thiol reductase needed for antigen presentation. (A) Transcript levels of *IFI30*, *LGMN*, and *CTSB* encoding GILT, AEP, and CTSB, respectively, were measured with quantitative RT-PCR ($n = 3$). P-values were calculated by a paired two-tailed Student's t test. *, $P < 0.05$. (B) HEK293T cells were cotransfected with different miRNA expression vectors and luciferase reporter plasmids carrying 3'-UTRs as indicated ($n = 3$). The luciferase activities were normalized to lysates from cells cotransfected with the wild-type 3'-UTR reporter and an empty plasmid. wt, wild-type 3'-UTR; mut, mutated 3'-UTR; \emptyset , empty plasmid. P-values were calculated by an unpaired two-tailed Student's t test. *, $P < 0.05$; ***, $P < 0.001$, with respect to the luciferase activity of the wild-type reporter cotransfected with an empty expression plasmid. (C) Transcript levels after RNAi knock-down of lysosomal enzyme-encoding genes were investigated. DG-75 cells were transfected with commercial siRNAs directed against *GAPDH*, *IFI30*, *LGMN*, or *CTSB* as indicated and transcript levels were quantified with RT-PCR. Means of three technical replicates are shown. (D) DG-75 cell transfected with siRNAs directed against three lysosomal enzymes as in (C) were loaded with purified influenza M1 protein and co-cultured with M1-specific CD4⁺ T cells (epitope LENV; HLA-DRB1*1301-restricted) for 1 d. After another 16 h, IFN- γ secretion was assessed by ELISA. One of two independent experiments is shown as means of three technical replicates.

2013). CD4⁺ T cells target a wide repertoire of EBV antigens from all phases of latent and lytic infection (Adhikary et al., 2007; Long et al., 2011). CD4⁺ T cells with specificity for structural EBV proteins play a prominent role: they are a universal component of the T cell repertoire, rapidly detecting EBV-infected B cells and killing them directly (Adhikary et al., 2006, 2007). Several EBV proteins expressed during its lytic phase can inhibit recognition of EBV-infected cells by CD4⁺ T cells (Ressing et al., 2015). The broad-ranging functions of the host shut-off gene product BGLF5 include reduction of HLA class II molecules on the cell surface during EBV's lytic, productive phase (Rowe et al., 2007). The late glycoprotein gp42, encoded by BZLF2, was shown to associate with HLA class II and to hinder recognition by CD4⁺ T cells sterically (Ressing et al., 2003). Both mechanisms are unlikely to be operational in newly infected B lymphocytes, because the two viral proteins appear not to be expressed in the prelatent phase (Kalla et al., 2010). Viral IL-10, encoded by BCLF1, is an immunomodulatory protein expressed early in infection, but its effects on B cell elimination by CD4⁺ T cells were limited (Jochum et al., 2012). Two additional viral

gene products reported to affect CD4⁺ T cell recognition, BDLF3 (Quinn et al., 2015) and BZLF1 (Zuo et al., 2011), may act during EBV primary infection, because BDLF3 is in the virus particle (Johannsen et al., 2004) and BZLF1 has been found to be expressed early during infection (Wen et al., 2007; Kalla et al., 2010). BDLF3 transcripts were present at very low levels, only, whereas BZLF1 transcripts were not mapped in our RNA-Seq analysis. Thus, how EBV infection escapes detection and elimination by EBV-specific T cells during the early phase of infection has remained uncertain.

Here, we present an answer to this question and show that EBV uses its large repertoire of miRNAs to target CD4⁺ T cell differentiation and recognition of infected cells. It appears that EBV's immunoevasive strategy uses miRNAs, which are themselves nonimmunogenic (Boss and Renne, 2011), rather than viral proteins which themselves would be antigenic. EBV induces a state of reduced immunogenicity in infected and recently activated B cells with viral miRNAs, which allows the virus to express its latency-associated antigens avoiding the recognition and elimination by CD4⁺ T cells. Because activated B cells are professional antigen-pre-

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

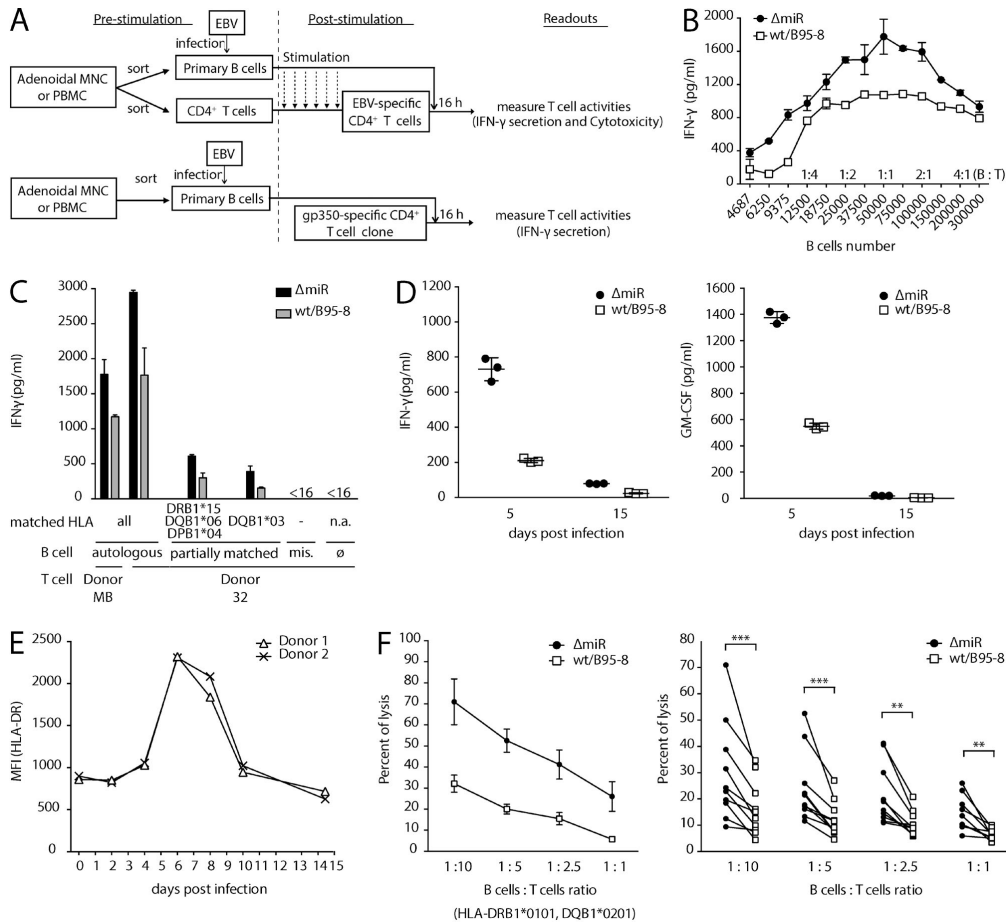


Figure 6. EBV miRNAs inhibit recognition and killing of infected B cells by EBV-specific CD4⁺ T cells. (A) Overview of the co-culture experiments used in B–F. Primary B cells sorted from adenoids or PBMCs were infected with either wt/B95-8 or ΔmiR EBV and co-cultured with polyclonal (top) or monoclonal (bottom) EBV-specific CD4⁺ T cells. Polyclonal antiviral CD4⁺ T cells were selected through stimulation (once in every two weeks) with irradiated LCLs infected with wt/B95-8 EBV. (B) Polyclonal EBV-specific CD4⁺ T cells were co-cultured for 16 h with autologous B cells that had been infected 5 d earlier. Levels of secreted IFN-γ were quantified by ELISA ($n = 3$). Several B/T cell ratios were used as indicated. Means \pm SD are shown. (C) Autologous, partially HLA-matched, or mismatched (mis.) B cells infected with wt/B95-8 or ΔmiR EBV ($n = 3$; Table S2) were cocultivated with polyclonal EBV-specific CD4⁺ T cells and secreted IFN-γ was quantified by ELISA after 16 h. The B/T cell ratio was 1:1. Matched HLA class II alleles are indicated. Means \pm SD are shown. <16 is under the detection limit (16 pg/ml); ∅, only T cells; n.a., not applicable. (D) The gp350-specific CD4⁺ T cell clone, epitope FGQ (HLA-DRB1*1301), was used as effector cells together with autologous B cells from donor JM (Table S2) as targets. B cells had been infected for 15 d with the two EBV strains as indicated and were used at an B/T cell ratio of 1:1. After 16 h of co-culture, levels of secreted IFN-γ and GM-CSF were quantified by ELISA. Means \pm SD are shown. (E) MFIs measured in flow cytometry analysis of cell surface HLA-DR in primary B cells infected with wt/B95-8 EBV from two different donors are shown. (F) Killing of EBV-infected B cells by EBV-specific CD4⁺ T cells was analyzed at various B/T cell ratios by Calcein release assays. A representative experiment with partially matched EBV-infected target B cells (left; $n = 3$) and a summary of all independent experiments with partially matched B cells (right) are shown. Paired samples from individual donors are connected by solid lines. Means \pm SD are shown. P-values were calculated by a paired two-tailed Student's *t* test. **, $P < 0.01$; ***, $P < 0.001$.

JEM

senting cells and express multiple immune-activating molecules (Wiesner et al., 2008), the need for EBV to control its host cell is more urgent than for other complex viruses that do not rely on professional antigen-presenting immune cells for their life-cycle.

In our experiments, recognition of early-stage infected B cells by CD4⁺ T cells was strongly inhibited by multiple mechanisms, pointing to the biological importance of the immunoevasive functions of viral miRNAs. First, *IL12B* is dramatically repressed in wt/B95-8 EBV-infected B cells compared with Δ miR EBV-infected B cells, leading to the down-regulation of three IL-12 family cytokines, IL-12B (IL-12p40), IL-12 (p35/p40), and IL-23 (p19/p40). At least five different viral miRNAs control the fate of the *IL12B* transcript, targeting multiple sites within its 3'-UTR, indicating its critical role in immune regulation by EBV-infected B cells and a redundancy or even cooperativity of viral miRNAs. Interestingly, the miRNAs controlling *IL12B* originate from different viral transcripts (Barth et al., 2011), suggesting a robust control of *IL12B* in all phases of EBV's life cycle and in the many cell types EBV infects, which show different patterns of miRNA expression (Cai et al., 2006; Qiu et al., 2011). Repression of *IL12B* may not only reduce CD4⁺ T cell differentiation, as shown here, but also regulate T cell effector functions (Curtis and Mescher, 2010).

Second, lysosomal proteolysis is regulated. Transcripts of lysosomal endopeptidases, AEP and CTSB, and a thiol reductase, GILT, which are involved in proteolytic degradation and HLA class II epitope generation (Blum et al., 2013), are direct targets of the miRNAs miR-BART1, miR-BART2, and miR-BHRF1-2. An siRNA-mediated knock-down of these three genes in human B cells reduced their recognition by clonal epitope-specific CD4⁺ T cells (Fig. 5, C and D; Milosevic et al., 2005) suggesting an important role of the three lysosomal enzymes in antigen processing and presentation via MHC class II molecules in human B cells.

Third, HLA class II surface levels are down-regulated in B cells infected with wt/B95-8 EBV (Fig. 4 B). We also tested if EBV miRNA targeted MHC class II molecules directly. miRNAs encoded in wt/B95-8 EBV failed to inhibit consistently four HLA-DRB1 alleles tested in dual luciferase reporter assays and also lacked functional miRNA-binding sites. Thus, the reduction of HLA class II molecules is likely indirect and may be a consequence of altered lysosomal processing of epitopes. Such phenomena might be particularly important early during infection when the EBV-activated B cells present antigenic peptides of virion components (Fig. 6 D) and the expression of HLA class II molecules peaks at the cell surface (Fig. 6 E). Together, these findings demonstrate that EBV miRNAs redundantly and robustly inhibit specific immune functions in newly infected B cells that may otherwise metabolize viral proteins into HLA class II-presented peptides for recognition by antiviral CD4⁺ T cells.

A limited number of observations have been made before on immunomodulatory functions of EBV miRNAs. The RNAs encoding the innate immune effector molecules

MICB (Nachmani et al., 2009) and *NLRP3* (Haneklaus et al., 2012) have been found in cell lines to be inhibited by EBV's miRNAs. They were not down-regulated in our experiments (Fig. 1 C, Table S1, and GSE75776), which examined their levels during early infection of primary B cells. Our studies with primary B cells likely avoided the adaptive changes that arise during long-term culturing in vitro.

It is not immediately apparent why B lymphocytes release proinflammatory cytokines upon infection with EBV. One explanation is based on two viral, noncoding RNAs, termed EBERs, which are contained in virions and are transcribed in all EBV-infected cells. EBERs are known to trigger the endosomal TLR3 receptor or the cytosolic RIG-I sensor signaling pathway and induce type I interferon and IL-6 synthesis in infected B cells (Samanta et al., 2006; Wu et al., 2007; Iwakiri et al., 2009), which might lead to the expression of proinflammatory cytokines found in higher concentrations in the supernatants of B cells infected with Δ miR than with wt/B95-8 EBV.

Interestingly, LMP1, a viral membrane protein that is predominantly expressed in the latent phase but also early upon B cell infection, activates the CD40 pathway, and induces IL-6, adhesion molecules, and important immune co-receptors (Kieser and Sterz, 2015). Several viral BART miRNAs have been reported to control LMP1 expression (Lo et al., 2007; Verhoeven et al., 2016). We confirmed that miR-BART16, which is not in the wt/B95-8 EBV strain, and miR-BART3, which was not previously known to target LMP1 directly, do target it (Fig. 4 F). In addition, miR-BART3 was recently reported to reduce LMP1 protein levels in HEK 293T cells (Verhoeven et al., 2016). We failed to identify direct targets among the many co-receptors and adhesion molecules down-regulated in cells infected with wt/B95-8 EBV (Fig. 4 C). It is possible that, on average, BART miRNAs repress LMP1 levels in cells infected with wt/B95-8 EBV, and thereby indirectly inhibit surface expression of some immune co-receptors and adhesion molecules, further reducing immune recognition of EBV-infected B lymphocytes.

EBV induces proliferation of the B cells it initially infects, and fosters their survival. We have found that EBV encodes miRNAs that regulate multiple facets of a host's adaptive immune response in newly infected B cells. EBV-infected B cells lacking viral miRNAs are deficient both in regulating these responses and in other miRNA-dependent functions, including an inhibition of apoptosis (Seto et al., 2010). These latter defects have precluded comparisons of B cells newly infected with wt/B95-8 or Δ miR in humanized mouse models. In infection experiments with these mice, we observed defects in persistence after infection with Δ miR EBV compared with wt/B95-8 EBV, possibly resulting from a combination of decreased survival and enhanced immune control (unpublished data; C. Münz, personal communication).

Collectively, our studies of a model that closely mimics physiological infection in the early phase show that EBV's miRNAs interfere with CD4⁺ T cell control through multiple mechanisms. They inhibit the secretion of cytokines,

inhibit antigen processing and presentation, inhibit the differentiation of CD4⁺ T cells, and counteract recognition and elimination of infected B cells by EBV-specific CD4⁺ effector T cells. The breadth of EBV's use of its miRNAs to inhibit adaptive immune responses is unprecedented and contributes to its efficient establishment of a lifelong infection.

MATERIALS AND METHODS

Patient samples

Surgically removed adenoids and PBMCs were obtained from anonymous patients and anonymous volunteer blood donors, respectively, from Munich, Germany. The use of this human material was approved by the local ethics committee (Ethikkommission bei der LMU München) in writing.

Separation of human primary cells

Human primary B and T cells were prepared from adenoidal mononuclear cells (MNCs) or PBMCs by Ficoll-Hypaque gradient centrifugation with Pancoll (PAN-Biotech). B cells, CD4⁺ T cells, CD8⁺ T cells, and naive CD4⁺ T cells were separated from adenoidal MNCs or PBMCs using MACS separator (Miltenyi Biotec) with CD19 MicroBeads, CD4 MicroBeads, CD8 MicroBeads, and Naive CD4⁺ T cell Isolation kit II, respectively.

Cell lines and cell culture

Burkitt's lymphoma cell lines Raji (EBV-positive), DG-75 (EBV-negative), HEK293-based EBV producer cell lines (Seto et al., 2010), infected human primary B cells, and T cells were maintained in RPMI-1640 medium (Thermo Fischer Scientific). HEK293T cells were maintained in DMEM medium. All media were supplemented with 10% FBS (Thermo Fischer Scientific), penicillin (100 U/ml; Thermo Fischer Scientific), and streptomycin (100 mg/ml; Thermo Fischer Scientific). Cells were cultivated at 37°C in a 5% CO₂ incubator.

Preparation of infectious EBV stocks and infection of human primary B cells

Infectious EBV stocks were prepared as previously described (Seto et al., 2010). In brief, EBV producer cell lines for ΔmiR (p4027) and wt/B95-8 (p2089) EBV strains were transiently transfected with expression plasmids encoding BZLF1 and BALF4 to induce EBV's lytic cycle. We collected supernatants 3 d after transfection, and debris was cleared by centrifugation at 3,000 rpm for 15 min. Virus stocks were titrated on Raji cells as previously reported and used at a multiplicity of infection (MOI) of 0.1 Green Raji units (Seto et al., 2010) for infecting primary B lymphocytes with an optimal virus dose (Steinbrück et al., 2015). For virus infection, primary B cells were cultivated with each virus stock for 18 h. After replacement with fresh medium, the infected cells were seeded at an initial density of 5 × 10⁵ cells per ml.

RNA-Seq and RISC-IP

At 5 dpi of human primary B cells, we extracted total RNAs with TRIzol (Thermo Fischer Scientific) and Di-

rect-Zol RNA MiniPrep kit (Zymo Research) from six different donors (Ad1-Ad6; Fig. 1) for RNA-Seq, according to the manufacturers' protocols. In parallel, we performed RISC immunoprecipitation (RISC-IP) as described previously (Kuzembayeva et al., 2012). In brief, lysed cells were incubated with anti-Ago2 antibody (11A9)-conjugated Dynabeads (Thermo Fischer Scientific), washed, and coprecipitated RNA was extracted. The cDNA libraries were prepared (Vertis Biotechnologie AG). For RNA-Seq, total RNAs were depleted of rRNAs by Ribo-Zero rRNA Removal kit (Illumina), fragmented by ultrasonication, and subjected to first strand synthesis with a randomized primer. For RISC-IP, RNAs were poly (A)-tailed, ligated with an RNA adapter at 5'-phosphates to facilitate Illumina TruSeq sequencing, and subjected to first strand synthesis with an oligo-(dT) primer. The cDNAs were PCR-amplified and sequenced with an Illumina HiSeq2000 instrument at the University of Wisconsin Biotechnology Center DNA Sequencing Facility.

Analysis of deep sequencing

For RNA-Seq, processing of paired-end reads (poly-A tail filtering, N-filtering, and adapter removal) was done using FastQC and R2M (RawReadManipulator). Reads were mapped to the human genome (hg19 'core' chromosome-set) by STAR (Dobin et al., 2013) and feature counts per transcript were determined using featureCounts and GENCODE version 19 annotations, together with EBV's annotation (available from GenBank under accession no. AJ507799). To screen differentially regulated genes by viral miRNAs, we used a simple but efficient scoring algorithm based on donor/replicate-wise fold changes ranks. For each gene *g* and replicate *k*, we calculate the gene-specific rank score as

$$r_g = \frac{1}{m} \sum_{k=1}^n r_{gk},$$

where *n* is the number of all replicates, *m* the number of all genes/transcripts, and *r_{gk}* is the rank of gene *g* in sample *k*. To select highly differentially expressed genes, we transformed the rank score into a z-score and selected all transcripts with an absolute z-score >1.6.

For RISC-IP the mapped reads were normalized using size factors estimated with the R package DESeq2 and filtered for reads mapped to annotated 3'-UTR regions using GENCODE version19. To identify local quantitative differences in the read enrichments on 3'-UTRs between wt EBV compared with ΔmiR EBV-infected B cells, we calculated a donor-wise relative enrichment score. For each genomic position *p*, the relative expression *es_p* was calculated as

$$es_p = \frac{e_p}{e_p + e_{cp}} \times n_{pus},$$

where *e_p* is the enrichment value of sequenced reads at position *p* in wt/B95-8 EBV-infected cells and *e_{cp}* the local enrichment value in ΔmiR EBV-infected B cells, respectively.

JEM

The normalization factor $n_{pu} = e_p / \max(e_u)$ was introduced to correct for local maxima in the UTR sequence of interest, where $\max(e_u)$ is the maximum enrichment value in the UTR sequence u . Finally, we used a Gaussian filter to minimize local noise. To select 3'-UTRs bound by viral miRNAs, we set the threshold as follows: enrichment score >0.6 for a stretch of >20 nt in the 3'-UTRs in two or more donors. To quantify viral miRNAs incorporated into the RISC in infected cells, we mapped reads from the RISC-IP Seq to miRNA entries registered in miRBase 21 and calculated fractions of each viral miRNAs out of total miRNA read counts.

KEGG enrichment pathway

Enrichment of specific pathways was estimated by performing a hypergeometric distribution test via the KEGG API Web Service. All calculations were done using Matlab (Mathworks).

ELISA

To detect cytokine secretion from infected B cells, 10^6 cells were seeded in 6-well plates at 4 or 11 dpi, cultivated for 4 d with cyclosporine (1 $\mu\text{g}/\text{ml}$; Novartis). Supernatants were harvested and stored at -20°C . ELISAs for IL-6, IL-10, IL12B (IL-12p40), IL-12, IL-23, and TNF were performed following the manufacturer's protocols (Mabtech). For IL-6 and IL-10, CpG DNA were added as previously described (Iskra et al., 2010) to stimulate infected B cells. ELISA for IFN- γ levels was performed following the manufacturer's protocol (Mabtech).

To detect IL-12p40 secretion from PBMCs or PBMCs depleted of B cells using the MACS separator and CD19 MicroBeads (Miltenyi Biotec), the cells were infected with either wt/B95-8 or ΔmiR EBV at MOIs of 0.1 Green Raji units (Steinbrück et al., 2015). After 5 d of incubation, supernatants were collected and ELISA for IL-12p40 levels was assessed following the manufacturer's protocol (Mabtech).

Luciferase reporter assays

The 3'-UTRs of *IL12B* (Ensembl ENST00000231228), *IFI30* (Ensembl ENST00000407280), *LGMN* (Ensembl ENST00000334869), *CTSB* (Ensembl ENST00000353047), and *LMP1* (available from GenBank under accession no. AJ507799) were cloned downstream of Renilla luciferase (*Rluc*) in the expression plasmid psiCHECK-2 (Promega). To construct the viral miRNA expression vectors, we cloned TagBFP (Evrogen) under the control of the EF1 α promoter into pCDH-EF1-MCS (System Biosciences). Single miRNAs of interest were cloned downstream of the TagBFP-encoding gene. Viral miRNAs were obtained by PCR from the p4080 plasmid (Seto et al., 2010). 50 ng of the psiCHECK-2 reporter and 150 ng of the pCDH-EF1 miRNA expressor plasmid DNAs were cotransfected into 1×10^5 HEK293T cells by Metafectene Pro (Biontex). After 24 h of transfection, we measured luciferase activities with the Dual-Luciferase Assay kit (Promega) and the Orion II Microplate Luminometer (Titertek-Berthold). The activity of Rluc was normal-

ized to the activity of Fluc (Firefly luciferase) encoded in the psiCHECK-2 reporter plasmid. We performed in silico prediction of EBV miRNA-binding sites on 3'-UTRs primarily with TargetScan (Garcia et al., 2011) and used RNAhybrid (Rehmsmeier et al., 2004) to screen for 6mer binding sites (Bartel, 2009). We performed site-directed mutagenesis with overlapping oligo DNAs and Phusion polymerase (NEB).

Quantitative RT-PCR

To quantify mRNA levels RNAs were reverse-transcribed with SuperScript III Reverse transcription (Thermo Fischer Scientific) and quantitative PCR was performed with LightCycler 480 SYBR Green I Mix (Roche) and LightCycler 480 Instrument II (Roche) according to the manufacturers' instructions. The following primers were used for the detection: *HPRT1* 5'-TGACCTTGATTTATTTGCATACC-3' and 5'-CGAGCAAGACGTTTCAGTCCT-3'; *HMBS* 5'-CTGAAAGGGCCTTCCTGAG-3' and 5'-CAGACTCCTCCA GTCAGGTACA-3'; *IL12B* 5'-CCCTGACATTCTGCCG TTCA-3' and 5'-AGGTCTTGTCCCGTGAAGACTCTA-3'; *IFI30* 5'-CTGGGTCACCGTCAATGG-3' and 5'-GCT TCTTGCCCTGGTACAAC-3'; *LGMN* 5'-GGAAAC TGATGAACACCAATGA-3' and 5'-GGAGACGATCTT ACGCACTGA-3'; *CTSB* 5'-CTGTGGCAGCATGTG TGG-3' and 5'-TCTTGTCCAGAAGTTCCAAGC-3'.

To quantify miRNA levels, stem-loop qPCRs were performed with TaqMan MicroRNA Reverse Transcription kit (Thermo Fischer Scientific) and TaqMan Universal Master Mix II (Thermo Fischer Scientific) according to the manufacturer's protocols. *RNU6B* was used for normalization. Following TaqMan MicroRNA assays, specific primers (Thermo Fischer Scientific) were used for detection: *ebv-miR-BART2*: 197238_mat; *ebv-miR-BART3*: 004578_mat; *ebv-miR-BHRF1-3* 197221_mat; *RNU6B*: 197238_mat.

Establishment of EBV-specific effector T cells and T cell clones

EBV-specific CD4⁺ T cell clones were established from polyclonal T cell lines that were generated by LCLs or mini-LCL stimulation of PBMCs, as previously described (Adhikary et al., 2007).

Flow cytometry and antibodies

After immunostainings with fluorophore-conjugated antibodies, single-cell suspensions were measured with LSR-Fortessa or FACSCanto (BD) flow cytometers and the FACSDiva software (BD). Acquired data were analyzed with FlowJo software Ver. 9.8 (FlowJo). The following fluorophore-conjugated antibodies reactive to human antigens were used: anti-human IFN- γ APC (4S.B3, IgG1; BioLegend), anti-CD40 PE (5c3, IgG2b; BioLegend), anti-ICOS-L (B7-H2) PE (2D3, IgG2b; BioLegend), anti-PD-L1 (B7-H1) APC (29E.2A3, IgG2b; BioLegend), anti-CD86 (B7-2) PE (37301, IgG1; R&D Systems), anti-CD54 (ICAM-1) APC (HCD54, IgG1; BioLegend), anti-HLA-ABC APC (W6/32, IgG2a;

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

BioLegend), anti-CD80 PE-Cy5 (L307.4; BD), anti-FAS (CD45) PE (Dx2, IgG1; BioLegend), anti-HLA-DR unlabeled (L234, IgG2a; BioLegend), anti-HLA-DQ unlabeled (SPV-L3, IgG2a ; AbD Serotec), anti-HLA-DP unlabeled (B7/21, IgG3; Abcam), anti-mouse F(ab')₂ APC (polyclonal, IgG; eBioscience), isotype IgG1 PE (MOPC-21; BioLegend), isotype IgG2b PE (MPC-11; BioLegend), isotype IgG1 APC (MOPC-21; BD), isotype IgG2a APC (MOPC-173; BioLegend), and isotype IgG2b APC (MG2b-57; BioLegend).

Western blotting

We lysed cells with RIPA buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% SDS, 1% NP-40, and 0.5% DOC) and boiled the extracts with Laemmli buffer. Proteins were separated on SDS-PAGE gels (Carl Roth) and transferred to nitrocellulose membranes (GE Healthcare) using Mini-PROTEAN Tetra Cell (Bio-Rad Laboratories). Membranes were blocked for 30 min with Roti-Block (Carl Roth), followed by antibody incubation. Secondary antibodies conjugated with horseradish peroxidase were used (Cell Signaling Technology) and exposed to CEA films (Agfa HealthCare). Protein levels were quantified with the software ImageJ. The following primary antibodies reactive to human proteins were used: anti-human Tubulin (B-5-1-2; Santa Cruz Biotechnology, Inc.). The monoclonal antibody (1G6-3) reactive to the EBV protein LMP1 was provided by E. Kremmer (Institute of Molecular Immunology, Helmholtz Zentrum München, München, Germany).

RNAi knock-down and recognition by M1-specific CD4⁺ T cells

4×10^5 DG-75 cells were incubated in 1 ml Accell Delivery Media (GE Healthcare) and 1 nmol siRNAs directed against *GAPDH*, *IFI30*, *LGMN*, *CTSB*, or combinations thereof for 48 h. Influenza M1 protein purified as previously described (Nimmerjahn et al., 2003) was added to the medium, and the cells were further incubated for 24 h and co-cultured with M1-specific CD4⁺ T cell clone E5 for 16 h (Milosevic et al., 2005). IFN- γ levels were detected with ELISA.

T cell differentiation and recognition

Th1 differentiation was assessed by co-culture of sorted naive CD4⁺ T cells and infected B cells 5 dpi. 1×10^5 naive CD4⁺ T cells stained with CellTrace Violet (Thermo Fischer Scientific) and 0.5 or 1×10^5 infected B cells were cultured in 96-well plates with Dynabeads Human T-Activator CD3/CD28 (Thermo Fischer Scientific) and cultivated for 7 d. The neutralizing antibody against IL12B (C8.6; BioLegend) or the corresponding isotype control antibody (MOPC-21; BioLegend) were added for certain experiments at 5 μ g/ml. Cells were restimulated with PMA and ionomycin (Cell Stimulation Cocktail; eBioscience) for 5 h and treated with Brefeldin A and Monensin (BioLegend) for 2.5 h before fixation. Th1 population was measured by intracellular IFN- γ staining with FIX and PERM Cell Fixation and Cell Permeabilization kit (Thermo Fischer Scientific) and subsequent flow cytometry

analysis. The Th1 population was defined as IFN- γ ⁺ T cells in the fraction of proliferating T cells identified via CellTrace Violet staining. EBV-specific effector T cells' activities were measured with ELISA and Calcein release assays. For IFN- γ detection from T cells, effector and target cells were seeded at 5×10^4 cell per ml (1:1 ratio) each and co-cultured for 16 h in a 96-well plate (V bottom). IFN- γ levels were detected with ELISA. IFN- γ concentrations <16 pg/ml were considered as not detected.

T cell cytotoxicity assays

Primary infected B cells were purified by Ficoll-Hypaque gradient centrifugation, and 5×10^5 target cells were labeled with calcein at 0.5 μ g/ml. After three washing steps with PBS, target and effector cells were co-cultured in a 96-well plate (V bottom) with different ratios in RPMI red phenol-free medium to reduce background signals. After 4 h of co-culture, fluorescence intensity of the released calcein was measured by the Infinite F200 PRO fluorometer (Tecan). As controls, spontaneous calcein release of target cells cultivated without effector cells and cells lysed with 0.5% Triton-X100 were used to define the levels of no and fully lysed target cells, respectively.

Statistical analysis

We used Prism 6.0 software (GraphPad) for the statistical analysis. A two-tailed ratio Student's *t* test was applied unless otherwise mentioned.

Online supplemental material

Fig. S1 shows the predicted miRNA-binding sites and mutations tested in 3'-UTR reporter assays. Table S1, available as an Excel file, lists the gene transcripts controlled by viral miRNAs. Table S2 lists the HLA allele information of donors used in co-culture experiments. Online supplemental material is available at <http://www.jem.org/cgi/content/full/jem.20160248/DC1>.

ACKNOWLEDGMENTS

We thank Christian Münz, Zurich, Elisabeth Kremmer, Anne-Wiebe Mohr, and Liridona Maliqi, Munich, for animal experiments, antibodies, T cell clones, and experimental assistance, respectively. We also thank Dagmar Pich for her experimental expertise and advice.

This work was financially supported by the Deutsche Forschungsgemeinschaft (SFB1054/TP B05 and TP A03, SFB1064/TP A13, and SFB-TR36/TP A04), Deutsche Krebshilfe (107277 and 109661), National Institutes of Health (R01: CA70723 and P01: CA022443), and personal grants of Deutscher Akademischer Austauschdienst to T. Tagawa (Studienstipendien für ausländische Graduierte aller wissenschaftlichen Fächer) and European Molecular Biology Organization to M. Bouvet.

The authors declare no competing financial interests.

Submitted: 18 February 2016

Accepted: 1 August 2016

JEM

REFERENCES

- Adhikary, D., U. Behrends, A. Moosmann, K. Witter, G.W. Bornkamm, and J. Mautner. 2006. Control of Epstein-Barr virus infection in vitro by T helper cells specific for virion glycoproteins. *J. Exp. Med.* 203:995–1006. <http://dx.doi.org/10.1084/jem.20051287>
- Adhikary, D., U. Behrends, H. Boerschmann, A. Pfänder, S. Burdach, A. Moosmann, K. Witter, G.W. Bornkamm, and J. Mautner. 2007. Immunodominance of lytic cycle antigens in Epstein-Barr virus-specific CD4⁺ T cell preparations for therapy. *PLoS One*. 2:e583. <http://dx.doi.org/10.1371/journal.pone.0000583>
- Bartel, D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 116:281–297. [http://dx.doi.org/10.1016/S0092-8674\(04\)00045-5](http://dx.doi.org/10.1016/S0092-8674(04)00045-5)
- Bartel, D.P. 2009. MicroRNAs: target recognition and regulatory functions. *Cell*. 136:215–233. <http://dx.doi.org/10.1016/j.cell.2009.01.002>
- Barth, S., G. Meister, and F.A. Grässer. 2011. EBV-encoded miRNAs. *Biochim. Biophys. Acta*. 1809:631–640. <http://dx.doi.org/10.1016/j.bbaggm.2011.05.010>
- Blum, J.S., P.A. Wearsch, and P. Cresswell. 2013. Pathways of antigen processing. *Annu. Rev. Immunol.* 31:443–473. <http://dx.doi.org/10.1146/annurev-immunol-032712-095910>
- Boss, I.W., and R. Renne. 2011. Viral miRNAs and immune evasion. *Biochim. Biophys. Acta*. 1809:708–714. <http://dx.doi.org/10.1016/j.bbaggm.2011.06.012>
- Cai, X., A. Schäfer, S. Lu, J.P. Bilello, R.C. Desrosiers, R. Edwards, N. Raab-Traub, and B.R. Cullen. 2006. Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathog.* 2:e23. <http://dx.doi.org/10.1371/journal.ppat.0020023>
- Curtsinger, J.M., and M.F. Mescher. 2010. Inflammatory cytokines as a third signal for T cell activation. *Curr. Opin. Immunol.* 22:333–340. <http://dx.doi.org/10.1016/j.coi.2010.02.013>
- Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 29:15–21. <http://dx.doi.org/10.1093/bioinformatics/bts635>
- Dölken, L., G. Malterer, F. Erhard, S. Kothe, C.C. Friedel, G. Suffert, L. Marciniowski, N. Motsch, S. Barth, M. Beitzinger, et al. 2010. Systematic analysis of viral and cellular microRNA targets in cells latently infected with human γ -herpesviruses by RISC immunoprecipitation assay. *Cell Host Microbe*. 7:324–334. <http://dx.doi.org/10.1016/j.chom.2010.03.008>
- Erhard, F., L. Dölken, L. Jaskiewicz, and R. Zimmer. 2013. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol.* 14:R79. <http://dx.doi.org/10.1186/gb-2013-14-7-r79>
- Feederle, R., J. Haar, K. Bernhardt, S.D. Linnstaedt, H. Bannert, H. Lips, B.R. Cullen, and H.-J. Delecluse. 2011a. The members of an Epstein-Barr virus microRNA cluster cooperate to transform B lymphocytes. *J. Virol.* 85:9801–9810. <http://dx.doi.org/10.1128/JVI.05100-11>
- Feederle, R., S.D. Linnstaedt, H. Bannert, H. Lips, M. Bencun, B.R. Cullen, and H.-J. Delecluse. 2011b. A viral microRNA cluster strongly potentiates the transforming properties of a human herpesvirus. *PLoS Pathog.* 7:e1001294. <http://dx.doi.org/10.1371/journal.ppat.1001294>
- Garcia, D.M., D. Baek, C. Shin, G.W. Bell, A. Grimson, and D.P. Bartel. 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat. Struct. Mol. Biol.* 18:1139–1146. <http://dx.doi.org/10.1038/nsmb.2115>
- Gottschalk, S., C.M. Rooney, and H.E. Heslop. 2005. Post-transplant lymphoproliferative disorders. *Annu. Rev. Med.* 56:29–44. <http://dx.doi.org/10.1146/annurev.med.56.082103.104727>
- Gottwein, E., D.L. Corcoran, N. Mukherjee, R.L. Skalsky, M. Hafner, J.D. Nusbaum, P. Shamulilatpam, C.L. Love, S.S. Dave, T. Tuschl, et al. 2011. Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe*. 10:515–526. <http://dx.doi.org/10.1016/j.chom.2011.09.012>
- Grundhoff, A., and C.S. Sullivan. 2011. Virus-encoded microRNAs. *Virology*. 411:325–343. <http://dx.doi.org/10.1016/j.virol.2011.01.002>
- Haneklaus, M., M. Gerlic, M. Kurowska-Stolarska, A.-A. Rainey, D. Pich, I.B. McInnes, W. Hammerschmidt, L.A.J. O’Neill, and S.L. Masters. 2012. Cutting edge: miR-223 and EBV miR-BART15 regulate the NLRP3 inflammasome and IL-1 β production. *J. Immunol.* 189:3795–3799. <http://dx.doi.org/10.4049/jimmunol.1200312>
- Haque, T., G.M. Wilkie, M.M. Jones, C.D. Higgins, G. Urquhart, P. Wingate, D. Burns, K. McAulay, M. Turner, C. Bellamy, et al. 2007. Allogeneic cytotoxic T-cell therapy for EBV-positive posttransplantation lymphoproliferative disease: results of a phase 2 multicenter clinical trial. *Blood*. 110:1123–1131. <http://dx.doi.org/10.1182/blood-2006-12-063008>
- Hislop, A.D., N.E. Anells, N.H. Gudgeon, A.M. Leese, and A.B. Rickinson. 2002. Epitope-specific evolution of human CD8(+) T cell responses from primary to persistent phases of Epstein-Barr virus infection. *J. Exp. Med.* 195:893–905. <http://dx.doi.org/10.1084/jem.20011692>
- IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. 2010. IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. *IARC Monogr. Eval. Carcinog. Risks Hum.* 94:v–vii: 1–412.
- Icheva, V., S. Kayser, D. Wolff, S. Tuve, C. Kyzirakos, W. Bethge, J. Greil, M.H. Albert, W. Schwinger, M. Nathrath, et al. 2013. Adoptive transfer of Epstein-Barr virus (EBV) nuclear antigen 1-specific T cells as treatment for EBV reactivation and lymphoproliferative disorders after allogeneic stem-cell transplantation. *J. Clin. Oncol.* 31:39–48. <http://dx.doi.org/10.1200/JCO.2011.39.8495>
- Iskra, S., M. Kalla, H.J. Delecluse, W. Hammerschmidt, and A. Moosmann. 2010. Toll-like receptor agonists synergistically increase proliferation and activation of B cells by Epstein-Barr virus. *J. Virol.* 84:3612–3623. <http://dx.doi.org/10.1128/JVI.01400-09>
- Iwakiri, D., L. Zhou, M. Samanta, M. Matsumoto, T. Ebihara, T. Seya, S. Imai, M. Fujieda, K. Kawa, and K. Takada. 2009. Epstein-Barr virus (EBV)-encoded small RNA is released from EBV-infected cells and activates signaling from Toll-like receptor 3. *J. Exp. Med.* 206:2091–2099. <http://dx.doi.org/10.1084/jem.20081761>
- Jochum, S., A. Moosmann, S. Lang, W. Hammerschmidt, and R. Zeidler. 2012. The EBV immunoevasins vIL-10 and BNLF2a protect newly infected B cells from immune recognition and elimination. *PLoS Pathog.* 8:e1002704. <http://dx.doi.org/10.1371/journal.ppat.1002704>
- Johannsen, E., M. Luffig, M.R. Chase, S. Weickel, E. Cahir-McFarland, D. Illanes, D. Sarracino, and E. Kieff. 2004. Proteins of purified Epstein-Barr virus. *Proc. Natl. Acad. Sci. USA*. 101:16286–16291. <http://dx.doi.org/10.1073/pnas.0407320101>
- Kalla, M., and W. Hammerschmidt. 2012. Human B cells on their route to latent infection—early but transient expression of lytic genes of Epstein-Barr virus. *Eur. J. Cell Biol.* 91:65–69. <http://dx.doi.org/10.1016/j.ejcb.2011.01.014>
- Kalla, M., A. Schmeinck, M. Bergbauer, D. Pich, and W. Hammerschmidt. 2010. AP-1 homolog BZLF1 of Epstein-Barr virus has two essential functions dependent on the epigenetic state of the viral genome. *Proc. Natl. Acad. Sci. USA*. 107:850–855. <http://dx.doi.org/10.1073/pnas.0911948107>
- Kieser, A., and K.R. Sterz. 2015. The Latent Membrane Protein 1 (LMP1). *Curr. Top. Microbiol. Immunol.* 391:119–149.
- Klinke, O., R. Feederle, and H.-J. Delecluse. 2014. Genetics of Epstein-Barr virus microRNAs. *Semin. Cancer Biol.* 26:52–59. <http://dx.doi.org/10.1016/j.semcancer.2014.02.002>
- Kuzembayeva, M., Y.-F. Chiu, and B. Sugden. 2012. Comparing proteomics and RISC immunoprecipitations to identify targets of Epstein-Barr

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

- viral miRNAs. *PLoS One*. 7:e47409. <http://dx.doi.org/10.1371/journal.pone.0047409>
- Lo, A.K.F., K.F. To, K.W. Lo, R.W.-M. Lung, J.W.Y. Hui, G. Liao, and S.D. Hayward. 2007. Modulation of LMP1 protein expression by EBV-encoded microRNAs. *Proc. Natl. Acad. Sci. USA*. 104:16164–16169. <http://dx.doi.org/10.1073/pnas.0702896104>
- Long, H.M., A.M. Leese, O.L. Chagoury, S.R. Connerty, J. Quarcoopome, L.L. Quinn, C. Shannon-Lowe, and A.B. Rickinson. 2011. Cytotoxic CD4⁺ T cell responses to EBV contrast with CD8 responses in breadth of lytic cycle antigen choice and in lytic cycle recognition. *J. Immunol*. 187:92–101. <http://dx.doi.org/10.4049/jimmunol.1100590>
- Milosevic, S., U. Behrends, H. Christoph, and J. Mautner. 2005. Direct mapping of MHC class II epitopes. *J. Immunol. Methods*. 306:28–39. <http://dx.doi.org/10.1016/j.jim.2005.07.020>
- Moosmann, A., I. Bigalke, J. Tischer, L. Schirrmann, J. Kasten, S. Tippmer, M. Leeping, D. Prevalsek, G. Jaeger, G. Ledderose, et al. 2010. Effective and long-term control of EBV PTLD after transfer of peptide-selected T cells. *Blood*. 115:2960–2970. <http://dx.doi.org/10.1182/blood-2009-08-236356>
- Nachmani, D., N. Stern-Ginossar, R. Sarid, and O. Mandelboim. 2009. Diverse herpesvirus microRNAs target the stress-induced immune ligand MICB to escape recognition by natural killer cells. *Cell Host Microbe*. 5:376–385. <http://dx.doi.org/10.1016/j.chom.2009.03.003>
- Nimmerjahn, F., D. Kobelt, A. Steinkasserer, A. Menke, G. Hobom, U. Behrends, G.W. Bornkamm, and J. Mautner. 2003. Efficient generation and expansion of antigen-specific CD4⁺ T cells by recombinant influenza viruses. *Eur. J. Immunol*. 33:3331–3341. <http://dx.doi.org/10.1002/eji.200324342>
- Qiu, J., K. Cosmopoulos, M. Pegtel, E. Hopmans, P. Murray, J. Middeldorp, M. Shapiro, and D.A. Thorley-Lawson. 2011. A novel persistence associated EBV miRNA expression profile is disrupted in neoplasia. *PLoS Pathog*. 7:e1002193. <http://dx.doi.org/10.1371/journal.ppat.1002193>
- Quinn, L.L., L.R. Williams, C. White, C. Forrest, J. Zuo, and M. Rowe. 2015. The missing link in Epstein-Barr virus immune evasion: the BDLF3 gene induces ubiquitination and downregulation of major histocompatibility complex class I (MHC-I) and MHC-II. *J. Virol*. 90:356–367. <http://dx.doi.org/10.1128/JVI.02183-15>
- Rancan, C., L. Schirrmann, C. Hüls, R. Zeidler, and A. Moosmann. 2015. Latent membrane protein LMP2A impairs recognition of EBV-infected cells by CD8⁺ T cells. *PLoS Pathog*. 11:e1004906. <http://dx.doi.org/10.1371/journal.ppat.1004906>
- Rehmsmeier, M., P. Steffen, M. Höchsmann, and R. Giegerich. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA*. 10:1507–1517. <http://dx.doi.org/10.1261/rna.5248604>
- Ressing, M.E., D. van Leeuwen, F.A.W. Verreck, R. Gomez, B. Heemskerck, M. Toebes, M.M. Mullen, T.S. Jardetzky, R. Longnecker, M.W. Schilham, et al. 2003. Interference with T cell receptor-HLA-DR interactions by Epstein-Barr virus gp42 results in reduced T helper cell recognition. *Proc. Natl. Acad. Sci. USA*. 100:11583–11588. <http://dx.doi.org/10.1073/pnas.2034960100>
- Ressing, M.E., D. Horst, B.D. Griffin, J. Tellam, J. Zuo, R. Khanna, M. Rowe, and E.J.H.J. Wiertz. 2008. Epstein-Barr virus evasion of CD8⁺ and CD4⁺ T cell immunity via concerted actions of multiple gene products. *Semin. Cancer Biol*. 18:397–408. <http://dx.doi.org/10.1016/j.semcancer.2008.10.008>
- Ressing, M.E., M. van Gent, A.M. Gram, M.J.G. Hooykaas, S.J. Piersma, and E.J.H.J. Wiertz. 2015. Immune evasion by Epstein-Barr virus. *Curr. Top. Microbiol. Immunol*. 391:355–381. http://dx.doi.org/10.1007/978-3-319-22834-1_12
- Riley, K.J., G.S. Rabinowitz, T.A. Yario, J.M. Luna, R.B. Darnell, and J.A. Steitz. 2012. EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO J*. 31:2207–2221. <http://dx.doi.org/10.1038/emboj.2012.63>
- Rowe, M., B. Glaunsinger, D. van Leeuwen, J. Zuo, D. Sweetman, D. Ganem, J. Middeldorp, E.J.H.J. Wiertz, and M.E. Ressing. 2007. Host shutoff during productive Epstein-Barr virus infection is mediated by BGLF5 and may contribute to immune evasion. *Proc. Natl. Acad. Sci. USA*. 104:3366–3371. <http://dx.doi.org/10.1073/pnas.0611128104>
- Samanta, M., D. Iwakiri, T. Kanda, T. Imaizumi, and K. Takada. 2006. EB virus-encoded RNAs are recognized by RIG-I and activate signaling to induce type I IFN. *EMBO J*. 25:4207–4214. <http://dx.doi.org/10.1038/sj.emboj.7601314>
- Seto, E., A. Moosmann, S. Grömminger, N. Walz, A. Grundhoff, and W. Hammerschmidt. 2010. Micro RNAs of Epstein-Barr virus promote cell cycle progression and prevent apoptosis of primary human B cells. *PLoS Pathog*. 6:e1001063. <http://dx.doi.org/10.1371/journal.ppat.1001063>
- Skalsky, R.L., and B.R. Cullen. 2010. Viruses, microRNAs, and host interactions. *Annu. Rev. Microbiol*. 64:123–141. <http://dx.doi.org/10.1146/annurev.micro.112408.134243>
- Skalsky, R.L., D.L. Corcoran, E. Gottwein, C.L. Frank, D. Kang, M. Hafner, J.D. Nusbaum, R. Feederle, H.-J. Delecluse, M.A. Luftig, et al. 2012. The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS Pathog*. 8:e1002484. <http://dx.doi.org/10.1371/journal.ppat.1002484>
- Steinbrück, L., M. Gustems, S. Medele, T.F. Schulz, D. Lutter, and W. Hammerschmidt. 2015. K1 and K15 of Kaposi's sarcoma-associated herpesvirus are partial functional homologues of latent membrane protein 2A of Epstein-Barr virus. *J. Virol*. 89:7248–7261. <http://dx.doi.org/10.1128/JVI.00839-15>
- Szabo, S.J., B.M. Sullivan, S.L. Peng, and L.H. Glimcher. 2003. Molecular mechanisms regulating Th1 immune responses. *Annu. Rev. Immunol*. 21:713–758. <http://dx.doi.org/10.1146/annurev.immunol.21.120601.140942>
- Thorley-Lawson, D.A. 2005. EBV the prototypical human tumor virus—just how bad is it? *J. Allergy Clin. Immunol*. 116:251–261. <http://dx.doi.org/10.1016/j.jaci.2005.05.038>
- Vereide, D.T., E. Seto, Y.-F. Chiu, M. Hayes, T. Tagawa, A. Grundhoff, W. Hammerschmidt, and B. Sugden. 2014. Epstein-Barr virus maintains lymphomas via its miRNAs. *Oncogene*. 33:1258–1264. <http://dx.doi.org/10.1038/onc.2013.71>
- Verhoeven, R.J.A., S. Tong, G. Zhang, J. Zong, Y. Chen, D.-Y. Jin, M.-R. Chen, J. Pan, and H. Chen. 2016. NF-κB signaling regulates expression of Epstein-Barr virus BART microRNAs and long noncoding RNAs in nasopharyngeal carcinoma. *J. Virol*. 00613–16. <http://dx.doi.org/10.1128/JVI.00613-16>
- Wen, W., D. Iwakiri, K. Yamamoto, S. Maruo, T. Kanda, and K. Takada. 2007. Epstein-Barr virus BZLF1 gene, a switch from latency to lytic infection, is expressed as an immediate-early gene after primary infection of B lymphocytes. *J. Virol*. 81:1037–1042. <http://dx.doi.org/10.1128/JVI.01416-06>
- Wiesner, M., C. Zentz, C. Mayr, R. Wimmer, W. Hammerschmidt, R. Zeidler, and A. Moosmann. 2008. Conditional immortalization of human B cells by CD40 ligation. *PLoS One*. 3:e1464. <http://dx.doi.org/10.1371/journal.pone.0001464>
- Wu, Y., S. Maruo, M. Yajima, T. Kanda, and K. Takada. 2007. Epstein-Barr virus (EBV)-encoded RNA 2 (EBER2) but not EBER1 plays a critical role in EBV-induced B-cell growth transformation. *J. Virol*. 81:11236–11245. <http://dx.doi.org/10.1128/JVI.00579-07>
- Xia, T., A. O'Hara, I. Araujo, J. Barreto, E. Carvalho, J.B. Sapucaia, J.C. Ramos, E. Luz, C. Pedroso, M. Manrique, et al. 2008. EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3. *Cancer Res*. 68:1436–1442. <http://dx.doi.org/10.1158/0008-5472.CAN-07-5126>

JEM

Zuo, J., A. Currin, B.D. Griffin, C. Shannon-Lowe, W.A. Thomas, M.E. Rensing, E.J.H.J. Wiertz, and M. Rowe. 2009. The Epstein-Barr virus G-protein-coupled receptor contributes to immune evasion by targeting MHC class I molecules for degradation. *PLoS Pathog.* 5:e1000255. <http://dx.doi.org/10.1371/journal.ppat.1000255>

Zuo, J., W.A. Thomas, T.A. Haigh, L. Fitzsimmons, H.M. Long, A.D. Hislop, G.S. Taylor, and M. Rowe. 2011. Epstein-Barr virus evades CD4⁺ T cell responses in lytic cycle through BZLF1-mediated downregulation of CD74 and the cooperation of vBcl-2. *PLoS Pathog.* 7:e1002455. <http://dx.doi.org/10.1371/journal.ppat.1002455>



Epstein–Barr virus microRNAs reduce immune surveillance by virus-specific CD8⁺ T cells

Manuel Albanese^{a,b,1}, Takanobu Tagawa^{a,b,1}, Mickaël Bouvet^{a,b}, Liridona Maliqi^{a,b}, Dominik Lutter^c, Jonathan Hoser^d, Maximilian Hastreiter^d, Mitch Hayes^e, Bill Sugden^e, Larissa Martin^{a,b}, Andreas Moosmann^{a,b}, and Wolfgang Hammerschmidt^{a,b,2}

^aResearch Unit Gene Vectors, German Research Center for Environmental Health, Helmholtz Zentrum München, D-81377 Munich, Germany; ^bGerman Centre for Infection Research (DZIF), D-81377 Munich, Germany; ^cInstitute of Diabetes and Obesity, German Research Center for Environmental Health, Helmholtz Zentrum München, D-85748 Garching, Germany; ^dInstitute of Bioinformatics and System Biology, German Research Center for Environmental Health, Helmholtz Zentrum München, D-85764 Neuherberg, Germany; and ^eMcArdle Laboratory for Cancer Research, University of Wisconsin–Madison, Madison, WI 53705

Edited by Thomas E. Shenk, Princeton University, Princeton, NJ, and approved August 19, 2016 (received for review April 21, 2016)

Infection with Epstein–Barr virus (EBV) affects most humans worldwide and persists life-long in the presence of robust virus-specific T-cell responses. In both immunocompromised and some immunocompetent people, EBV causes several cancers and lymphoproliferative diseases. EBV transforms B cells in vitro and encodes at least 44 microRNAs (miRNAs), most of which are expressed in EBV-transformed B cells, but their functions are largely unknown. Recently, we showed that EBV miRNAs inhibit CD4⁺ T-cell responses to infected B cells by targeting IL-12, MHC class II, and lysosomal proteases. Here we investigated whether EBV miRNAs also counteract surveillance by CD8⁺ T cells. We have found that EBV miRNAs strongly inhibit recognition and killing of infected B cells by EBV-specific CD8⁺ T cells through multiple mechanisms. EBV miRNAs directly target the peptide transporter subunit TAP2 and reduce levels of the TAP1 subunit, MHC class I molecules, and EBNA1, a protein expressed in most forms of EBV latency and a target of EBV-specific CD8⁺ T cells. Moreover, miRNA-mediated down-regulation of the cytokine IL-12 decreases the recognition of infected cells by EBV-specific CD8⁺ T cells. Thus, EBV miRNAs use multiple, distinct pathways, allowing the virus to evade surveillance not only by CD4⁺ but also by antiviral CD8⁺ T cells.

adaptive immunity | immune evasion | herpesvirus | CD8 T cells | microRNA

Epstein–Barr virus (EBV) is a ubiquitous herpesvirus that infects the majority of the human population worldwide. Although EBV infection persists for life, most carriers remain asymptomatic due to a stringent control by virus-specific immunity. An important component of this immunity is EBV-specific CD8⁺ T cells, which often expand to high numbers in healthy carriers or after primary infection. Conversely, the absence of EBV-specific CD8⁺ T cells predicts the emergence of EBV-associated disease in patients after stem cell transplantation or when afflicted with AIDS (1–3). Dangerous EBV-mediated complications can be reversed or prevented by transfer of EBV-specific T cells (4, 5), which further confirms the important role of continuous T-cell control of EBV infection. Among EBV-specific T cells, CD8⁺ T cells predominate; about 0.05–1% of all CD8⁺ T cells in healthy donors are typically specific for EBV latent antigens and about twice as many for lytic antigens (6, 7).

EBV predominantly infects B cells and establishes a latent infection before production of progeny virus becomes possible (8). Four distinct programs of EBV latent infection have been defined according to their expression profiles of latent viral genes (9–11). One of these programs, known as latency III or the “growth program,” is characterized by the expression of a restricted set of approximately eight viral proteins, which activate B cells and drive their proliferation, thus increasing the viral reservoir. Latency III is found in EBV-associated malignancies in immunosuppressed patients (9) and likely reemerges continuously in healthy carriers (9, 12), indicating that its control is critical for the health of an EBV carrier. Only at a later stage of infection (13) can the virus

enter its lytic phase in infected cells to produce progeny virions, a phase requiring expression of the majority of viral proteins. Some of these lytic-cycle viral proteins are immunoevasins that interfere with CD8⁺ T-cell recognition: the TAP inhibitor BNLF2a (14, 15); the G protein-coupled receptor BILF1 that associates with MHC class I/peptide complexes, diverts them from the exocytic pathway and the cell surface, and induces their lysosomal degradation (16, 17); and the protein BGLF5 that reduces MHC I expression and CD8⁺ T-cell recognition as a consequence of its generalized host-shutoff function (18, 19). Recently, BDLF3 was identified as an additional lytic-cycle protein that targets MHC molecules for degradation (20). BNLF2a is also expressed early after infection in the prelatency phase (13 for a recent review) and reduces CD8⁺ T-cell recognition in the first days of infection but does not impair T-cell recognition in established latency (21, 22). It is unclear, though, how latently EBV-infected B cells escape elimination by T cells, in particular during the latency III program that is characterized by a considerable antigenic load and an activated state expected to increase the immunogenicity of B cells (23). In latency III, MHC I, MHC II, and T-cell-coactivating molecules are highly expressed (24, 25), but nonetheless many epitopes are suboptimally recognized by CD8⁺ T cells (26, 27). Therefore, it is likely that unknown immunoevasive mechanisms operate in these latently infected B cells.

Significance

Most humans are infected for their lifetime with Epstein–Barr virus (EBV), which can cause cancer and other EBV-associated diseases. Infected individuals develop strong immune responses to this virus, in particular cytotoxic CD8⁺ T cells, but viral infection is never cleared nor is EBV eliminated from the body. This suggests that certain viral molecules might prevent effective elimination of EBV-infected cells by CD8⁺ T cells. EBV is rich in genes coding for microRNAs, many with unknown function. We show that viral microRNAs interfere with recognition and killing of EBV-infected cells by CD8⁺ T cells. Multiple mechanisms and molecules are targeted by microRNAs to achieve this immune evasion. Therefore, targeting of viral microRNAs may improve antiviral immunity and therapy.

Author contributions: M.A., T.T., A.M., and W.H. designed research; M.A., T.T., M.B., and L. Maliqi performed research; L. Martin contributed new reagents/analytic tools; M.A., T.T., D.L., J.H., M. Hastreiter, and M. Hayes analyzed data; and M.A., T.T., B.S., A.M., and W.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹M.A. and T.T. contributed equally to this work.

²To whom correspondence should be addressed. Email: hammerschmidt@helmholtz-muenchen.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1605884113/-DCSupplemental.

A hallmark of EBV is its array of 44 microRNAs (miRNAs) (28–31), which is the largest number of miRNAs identified in a human pathogen to date. Many EBV miRNAs have no known function. A function of viral miRNAs in innate immunity was suggested earlier by findings showing that some regulate the inflammation component NLRP3 (32), the natural killer group 2D (NKG2D) ligand MICB (33), and the chemokine CXCL11 (34).

Recently, we found that multiple viral miRNAs limit the control of infected B cells by CD4⁺ T cells early in EBV infection (35). Several viral miRNAs reduce secretion of IL-12, expression of HLA class II molecules, and expression of lysosomal enzymes important for antigen presentation to CD4⁺ T cells. EBV miRNAs also regulate many molecules of potential importance in HLA class I presentation and CD8⁺ T-cell recognition (35).

These findings have led us to determine if viral miRNAs inhibit surveillance of EBV by CD8⁺ T cells. We have found that viral miRNAs do prevent virus-specific activation of CD8⁺ T cells and killing of infected B cells; we have also delineated the mechanisms underlying this viral evasion of the immune response.

Results

EBV miRNAs Support Survival of Infected B Cells in the Presence of CD8⁺ T Cells from EBV-Positive Donors. We used in vitro infection of primary human B cells as a simple but representative model of EBV infection (Fig. 1A) to evaluate if EBV miRNAs counteract immune surveillance by EBV-specific CD8⁺ T cells. B cells (CD19⁺) isolated from peripheral blood mononuclear cells (PBMCs) from EBV-positive donors were infected with two EBV strains: the laboratory strain B95-8 (WT/B95-8), expressing 13 viral miRNAs, and its derivative Δ miR, expressing no viral miRNAs (36). Twelve hours later, autologous CD8⁺ T cells were added, and the cells were cocultured for 4 wk (Fig. 1B), when cell viability was tested in MTT assays. Surviving cells were further analyzed by flow cytometry for their identification (Fig. S1).

In the absence of T cells, we observed robust proliferation of B cells infected with Δ miR or WT/B95-8 EBV (Fig. 1B). In the presence of T cells, the survival and outgrowth of infected B cells was decreased. A strong reduction of viable cells was achieved by fewer CD8⁺ T cells for cells infected with Δ miR EBV than cells infected with WT/B95-8 EBV (Fig. 1B). Flow cytometry analyses showed that B cells represented the viable cells in most of the cultures (Fig. S1). These results indicated that EBV miRNAs protect EBV-infected B lymphocytes from eradication by antiviral CD8⁺ T cells under these conditions.

EBV miRNAs Inhibit Recognition, Killing, and Expansion of EBV-Specific CD8⁺ T Cells. Earlier studies had shown that B-cell survival could be compromised in cells infected with the Δ miR EBV devoid of miRNAs, because some viral miRNAs contribute to EBV-associated cellular transformation in the early phase of infection (36–38). To evaluate a possible role of viral miRNAs in controlling immune functions of EBV-specific CD8⁺ T cells, we established polyclonal EBV-specific CD8⁺ T cells from different donors. Sorted primary CD8⁺ T cells were stimulated every 2 wk with irradiated lymphoblastoid cell lines (LCLs), which had been established by infecting primary autologous B cells with WT/B95-8 EBV (Fig. 2A). For T-cell effector assays, EBV-specific CD8⁺ T cells established in this way were cocultured with B cells that had been infected with WT/B95-8 or Δ miR EBV 15 d earlier. T-cell activation was quantified by measuring IFN- γ concentration in the cell culture supernatants after 16 h or by determining cytolysis of infected cells after 4 h. We observed significantly reduced IFN- γ secretion in response to cells infected with WT/B95-8 relative to cells infected with Δ miR EBV, both in autologous (Fig. 2B and Table S1) and HLA-matched conditions (Fig. 2C). Importantly, T cells were not activated by HLA-mismatched infected B cells or in B-cell-free cultures, indicating that the observed activation was HLA-restricted and

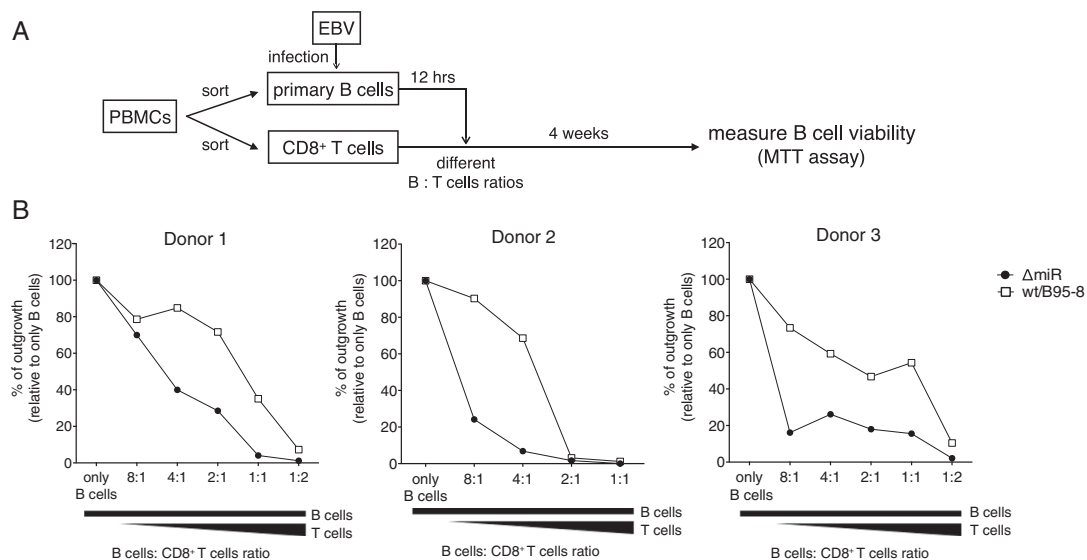


Fig. 1. EBV miRNAs support infected B cells to abrogate CD8⁺ T-cell responses. (A) Schematic overview of the experimental system. (B) CD19⁺ B cells were isolated from PBMCs of three different EBV-positive donors and infected with WT/B95-8 EBV or Δ miR EBV stocks. Twelve hours later, infected B cells were extensively washed to remove free virions. In 96-well microtiter plates 32,000 EBV-infected B cells were seeded per well and CD8⁺ T cells isolated from the autologous donors were added at different ratios as indicated. After 4 wk, total cell viability was assessed by MTT assay. Outgrowth of B-cell-only conditions (without T cells) was set to 100%. The results shown are based on the mean of six technical replicates per data point.

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

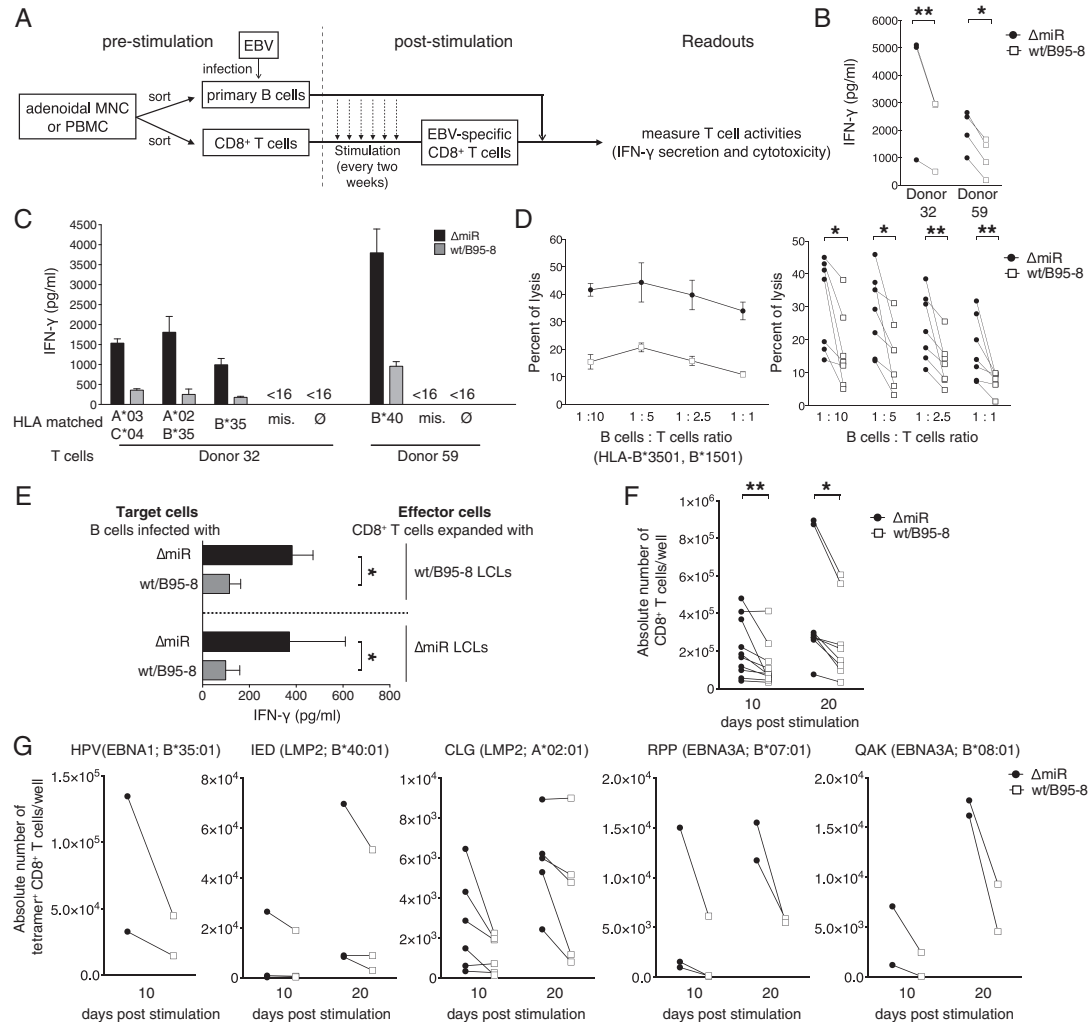


Fig. 2. EBV miRNAs inhibit recognition and killing of infected B cells as well as expansion of EBV-specific CD8⁺ T cells. (A) Schematic overview of the experiments shown in the remaining panels of this figure. Polyclonal EBV-specific CD8⁺ T cells were obtained by repeated stimulation (every two weeks) with autologous irradiated WT/B95-8 EBV-infected LCLs. The T-cell activities of the EBV-stimulated CD8⁺ T cell were subsequently analyzed with target B cells infected with WT/B95-8 EBV or ΔmiR EBV stocks. (B) Equal numbers of polyclonal EBV-specific CD8⁺ T cells and autologous B cells infected for 15 d with the indicated EBV strains were cocultured. After 16 h, IFN-γ released from T cells was measured by ELISA. Results of three to four biological replicates are shown for each donor. (C) Polyclonal EBV-specific CD8⁺ T cells were cocultured with HLA-matched or mismatched EBV-infected B cells and tested as in B. Matched HLA class I alleles are indicated; mis., mismatched; ∅, only T cells; <16, below the threshold of detection (16 pg/mL). HLA allotypes of the donors are listed in Table S1. Data are shown as mean values. Error bars indicate SD of three replicates. (D) Cytotoxic activities of EBV-specific CD8⁺ T cells directed against HLA-matched infected B cells were analyzed at various B:T cells ratios in calcein release assays after 4 h of coculture as shown in A. A representative experiment with mean values and SD of four replicates (Left) and the overview of seven donors (Right) are shown. (E) CD8⁺ T cells of donor 115 were repetitively stimulated for 2 mo with irradiated autologous B cells infected with WT/B95-8 or ΔmiR EBV as indicated. The expanded effector T cells were assayed with autologous infected B cells as targets in coculture experiments and tested as in B measuring the IFN-γ released by ELISA. Error bars indicate SD of three biological replicates. (F and G) CD8⁺ T cells isolated from PBMCs were stimulated on day 0 and 10 with irradiated B cells infected with the indicated EBV strains. Absolute numbers of T cells were determined by flow cytometry 10 d later (on day 10 and 20). (F) Expansion of total CD3⁺ CD8⁺ T cells. (G) Expansion of EBV epitope-specific CD8⁺ T cells, stained with HLA/peptide pentamers as indicated. Results from 3 to 10 different donors are shown. The significance of difference in the T-cell expansion experiments was calculated by Wilcoxon matched-pairs signed rank test. **P* < 0.05, ***P* < 0.01.

EBV-specific (Fig. 2C). In cytotoxicity assays, we found that the viral miRNAs inhibited killing of infected B cells by EBV-specific CD8⁺ T cells at all B:T cell ratios and at any HLA-matched

conditions tested (Fig. 2D). It did not matter whether EBV-specific CD8⁺ T-cell cultures had been generated by expansion with ΔmiR or WT/B95-8 EBV-infected B cells; in each case, ΔmiR

EBV-infected B cells were better recognized than B cells infected with WT/B95-8 EBV (Fig. 2E).

Because clonal expansion is essential for effective antiviral T-cell responses, we also investigated the selective expansion of EBV-specific CD8⁺ T cells in response to autologous B cells infected with either WT/B95-8 or ΔmiR EBV (Fig. 2F). From PBMCs of 10 donors, we sorted CD8⁺ T cells and stimulated them twice on day 0 and day 10 with irradiated autologous B cells, which had been infected for 15 d with WT/B95-8 or ΔmiR EBV. Total numbers of CD8⁺ T cells obtained at days 10 and 20 (after one or two stimulations) were significantly higher after expansion with ΔmiR EBV-infected cells compared with WT/B95-8 EBV-infected cells (Fig. 2F). In the same setting, we also analyzed the expansion of EBV-specific CD8⁺ T cells that were specific for five different epitopes from EBV proteins latent membrane protein (LMP)2A, EBV nuclear antigen (EBNA)1, and EBNA3A (Fig. 2G). We consistently found increased expansion in response to ΔmiR EBV-infected cells for each of these specificities (Fig. 2G). Together, our data suggest that viral miRNAs in EBV-infected B cells reduce clonal expansion of a wide range of antiviral CD8⁺ effector T cells.

EBV miRNAs Inhibit MHC Class I Antigen Processing and Presentation Pathways. We screened cellular transcripts targeted by EBV miRNAs and likely critical in fending off antiviral CD8⁺ T cells. To identify potential targets, we performed high-throughput screening with primary B lymphocytes infected with the different EBV strains and a combination of RNA and RNA induced silencing complexes-immunoprecipitation (RISC-IP) sequencing (35). With this approach, we identified *IL12B* and three genes (*IFI30*, the IFN-γ-regulated thiol reductase *GILT*; *LGGMN*, the asparagine endopeptidase *AEP* alias legumain; and *CTSB*, the peptidase cathepsin B) encoding lysosomal enzymes and important for CD4⁺ T-cell differentiation and antigen processing as direct targets of viral miRNAs (35). Here, we focused on genes consistently inhibited by EBV miRNAs and known to play a role in antigen processing and presentation and cytokine–cytokine receptor interactions or are considered cell adhesion molecules according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway categories (Table 1). A subset of corresponding mRNAs was also enriched in our RISC-IP sequencing analysis (Table 1), indicating that these mRNAs were most likely direct targets of EBV miRNAs. Interestingly, *TAP1* and *TAP2* were significantly down-regulated in RNA-sequencing (RNA-Seq) experiments, and *TAP2* was also found enriched in RISC-IP sequencing (Table 1). The TAP1/TAP2 heterodimer mediates transport of antigenic peptides into the ER lumen, where they are loaded onto MHC class I molecules. The presentation of many EBV epitopes depends on TAP (39), and thus we delineated the mechanisms by which EBV miRNAs regulate it.

First, we verified the regulation of *TAP1/2* expression by viral miRNAs. Fifteen days post infection expression of *TAP1* and *TAP2* was reduced in B cells infected with WT/B95-8 EBV compared with ΔmiR EBV both at the level of transcript (Fig. 3A) and protein (Fig. 3B). As a control, we verified that *IPO7* (Importin-7), a known target of EBV miR-BART3 (40), was also down-regulated (Fig. 3A and B). Because RISC-IP (Table 1) in combination with the in silico target algorithm TargetScan (41) predicted that the 3'UTR of *TAP2* was directly targeted by EBV miRNAs, we performed dual luciferase reporter assays to test this assumption. We cotransfected HEK293T cells with a luciferase reporter plasmid containing the 3'UTR of *TAP2* and single expression plasmids, each of which encoded one viral primary miRNA.

The expression of exogenous miR-BHRF1-3 significantly decreased the luciferase activity of the *TAP2* reporter (Fig. 3C, Left). A mutation within the 3'UTR in the seed-matching region abolished this inhibition completely, demonstrating that *TAP2* is a direct target of miR-BHRF1-3 (Fig. 3C). Similarly, miR-BART17, which is expressed in EBV field strains but not in the WT/B95-8

Table 1. Selected genes (cytokine–cytokine receptor interaction, antigen processing and presentation, and cell adhesion molecules) and their regulation by EBV miRNAs in RNA-Seq and RISC-IP experiments

Gene symbol	Mean of log-two-fold change	z score	RISC-IP
IL12B	−2.3108	−2.5002	Yes
CD274	−0.6644	−2.3705	No
CD80	−0.9250	−2.3533	No
ICAM1	−0.6553	−2.3351	No
TAP2	−0.4012	−2.1261	Yes
TNF	−0.6389	−2.0936	No
CD40	−0.4912	−2.0739	Yes
CD58	−0.5088	−2.0568	No
RFX5	−0.3576	−1.9605	No
PSME1	−0.3338	−1.9373	No
CTSB	−0.4520	−1.9370	Yes
PSME2	−0.4680	−1.9335	No
CD86	−0.4486	−1.8001	No
TAP1	−0.3426	−1.7800	No
ALCAM	−0.4300	−1.6637	No
ICAM3	−0.5567	−1.6273	No
ERAP2	−0.2923	−1.6048	No
IPO7	−0.2826	−1.6870	Yes

Genes were identified by mRNA sequencing of WT/B95-8 vs. ΔmiR-infected B cells, ranked by z score, and where indicated, confirmed by RISC-IPs as described in Tagawa and coworkers (35). *IPO7* served as a positive control.

strain, targeted the 3'UTR of the *TAP2* directly at one of two predicted sites (Fig. 3C; predicted seed sequences are provided in Fig. S2A). In contrast, we did not observe a regulation of the *TAP1* 3'UTR by any viral miRNA present in WT/B95-8 EBV (Fig. 3D). This result suggested that *TAP1* may not be a direct target of EBV miRNAs, consistent with our RISC-IP data (Table 1). A parallel dual luciferase reporter assay performed for *IPO7* served as a positive control together with miR-BART3 in these assays (Fig. S2B).

Next, we quantified the levels of classical HLA class I (HLA-A, -B, and -C) cell-surface expression on WT/B95-8 or ΔmiR EBV-infected B cells during the course of infection. Steady-state surface levels of HLA class I molecules are a function of TAP activities, as HLA class I molecules lacking peptides are unstable. We consistently observed a slight reduction by 10–20% of overall surface MHC class I molecules in cells infected with WT/B95-8 relative to ΔmiR EBV during the entire observation period (Fig. 3E). By assaying individual HLA class I alleles, we found that HLA-B*07, B*08, and B*40 allotypes were reduced by 20–30%, whereas HLA-A*02 levels were not reduced (Fig. 3F). This finding is consistent with the known preference of HLA-A*02 (but not of the other allotypes investigated here) to bind highly hydrophobic peptides, some of which reach the ER independently of TAP (42). Dual luciferase reporter assays were performed for HLA-B*07 and B*08, but direct targeting by miRNAs could not be demonstrated (Fig. S2C).

EBV miRNAs Control Multiple Facets of Viral Immune Evasion. These results suggested that EBV miRNAs impose allele-specific controls of HLA molecules, namely affecting HLA-B allotypes. We therefore asked if HLA-B allotype-restricted antigen presentation is directly controlled by viral miRNAs (Fig. 4). We cocultured infected B cells and CD8⁺ T-cell clones specific for the IED or the FLY epitope, both of which are derived from the viral LMP2 protein (Fig. 4A). Presentation of the B*40:01-restricted IED epitope is dependent on active TAP transportation (39), whereas the HLA-A*02:01-restricted epitope FLY is highly hydrophobic and presented TAP-independently (26). In a time course experiment

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

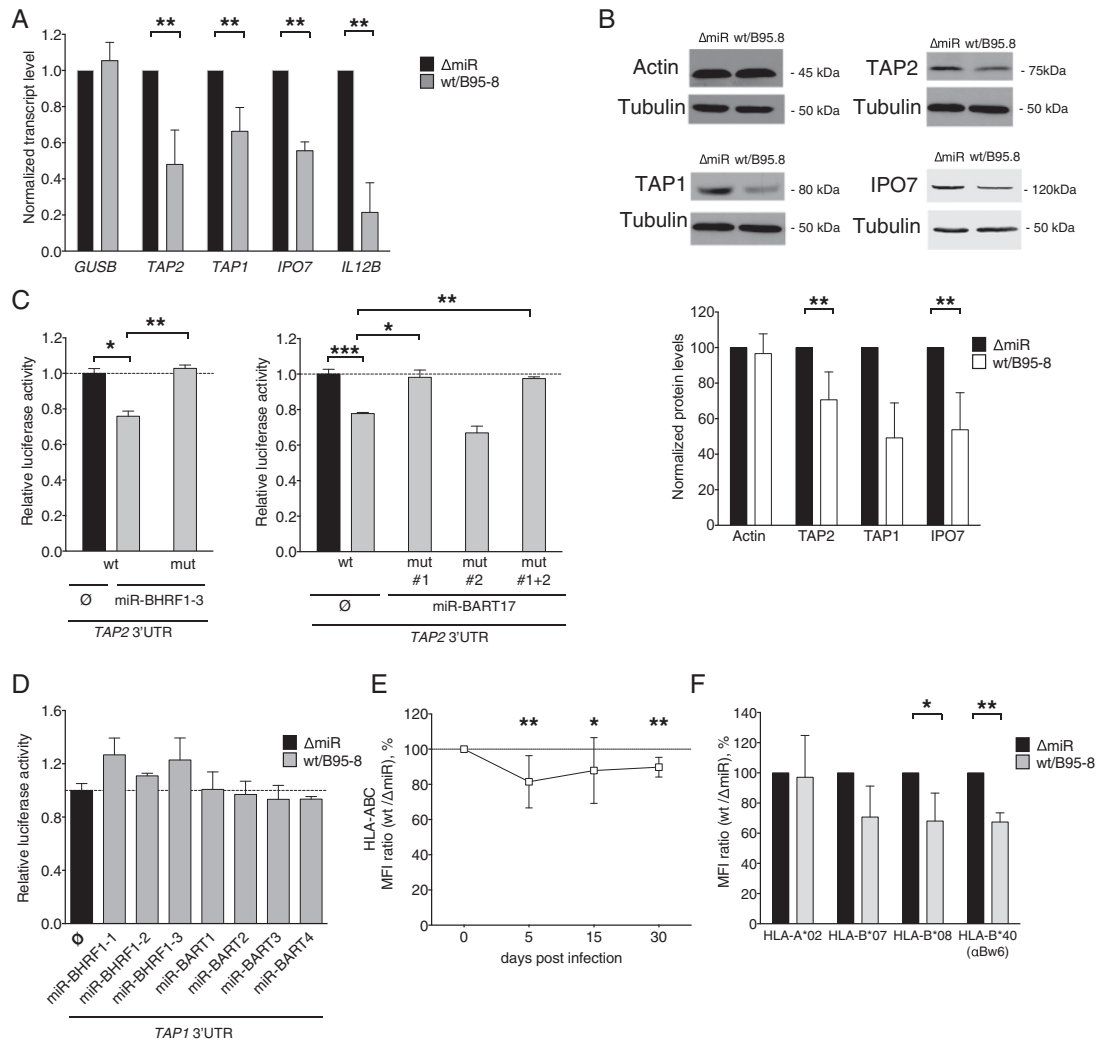


Fig. 3. EBV miRNAs reduce TAP and MHC class I levels in infected B cells. (A) Transcript levels of *TAP2*, *TAP1*, *IPO7*, and *IL12B* were assessed by quantitative RT-PCR in EBV-infected B cells 15 d post infection (dpi). *IPO7* is a known target of viral miRNAs and is used here as positive control. *GUSB* was used as negative control. Transcript levels were quantified relative to the mean of the housekeeping genes *HPRT1* and *HMBS* (35) and were normalized to the transcript level of Δ miR EBV-infected cells. Data are shown as mean values and SD of seven donors. (B) Protein levels of TAP1 and TAP2 were assessed by Western blot analyses in EBV-infected B cells 15 dpi. β -Actin served as negative and IPO7 as positive controls. Representative examples (Top) and protein levels relative to Tubulin (Bottom) are shown. The results were normalized to the protein levels of Δ miR EBV-infected cells, set to 100%. Data are shown as mean values and SD of three to seven donors. (C and D) EBV miRNAs directly regulate *TAP2* but not *TAP1*. HEK293T cells were cotransfected with miRNA expression vectors and dual luciferase reporter plasmids carrying a wild-type or mutated 3'UTR of *TAP2* (C). For the analysis of *TAP1* (D), all viral miRNAs present in WT/B95-8 EBV were tested with the exception of miR-BART15, which is barely expressed in our infection model (35). Sequence details of the 3'UTRs are contained in Fig. S2. The luciferase activities were normalized to lysates from cells cotransfected with the wild-type 3'UTR reporter and an empty plasmid in place of the miRNA expression plasmid. Data are shown as mean values and SD of three to four replicates. mut, mutated 3'UTR; WT, wild-type 3'UTR; \emptyset , empty plasmid. (E and F) Cell surface expression levels of total HLA class I (Left) and specific HLA class I allotypes (Right) of B cells infected with the indicated EBV strains for 15 d were measured by flow cytometry. Ratios (%) of WT/B95-8 divided by Δ miR EBV-infected B cells are shown. Data are shown as mean values and SD of experiments with 5 to 10 different donors. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

with B cells infected with either WT/B95-8 or Δ miR EBV, we observed reactivity of the clonal T cells as early as 5–7 d post infection (Fig. 4 B and C). EBV miRNAs significantly reduced the activation of the IED-specific T-cell clone (Fig. 4B) as expected from the

down-regulation of the TAP complex and subsequent reduction of HLA-B molecules. Surprisingly, the activation of the FLY-specific T-cell clone was also strongly reduced (Fig. 4C) even though FLY is a TAP-independent peptide and presented via HLA-A*02:01, which

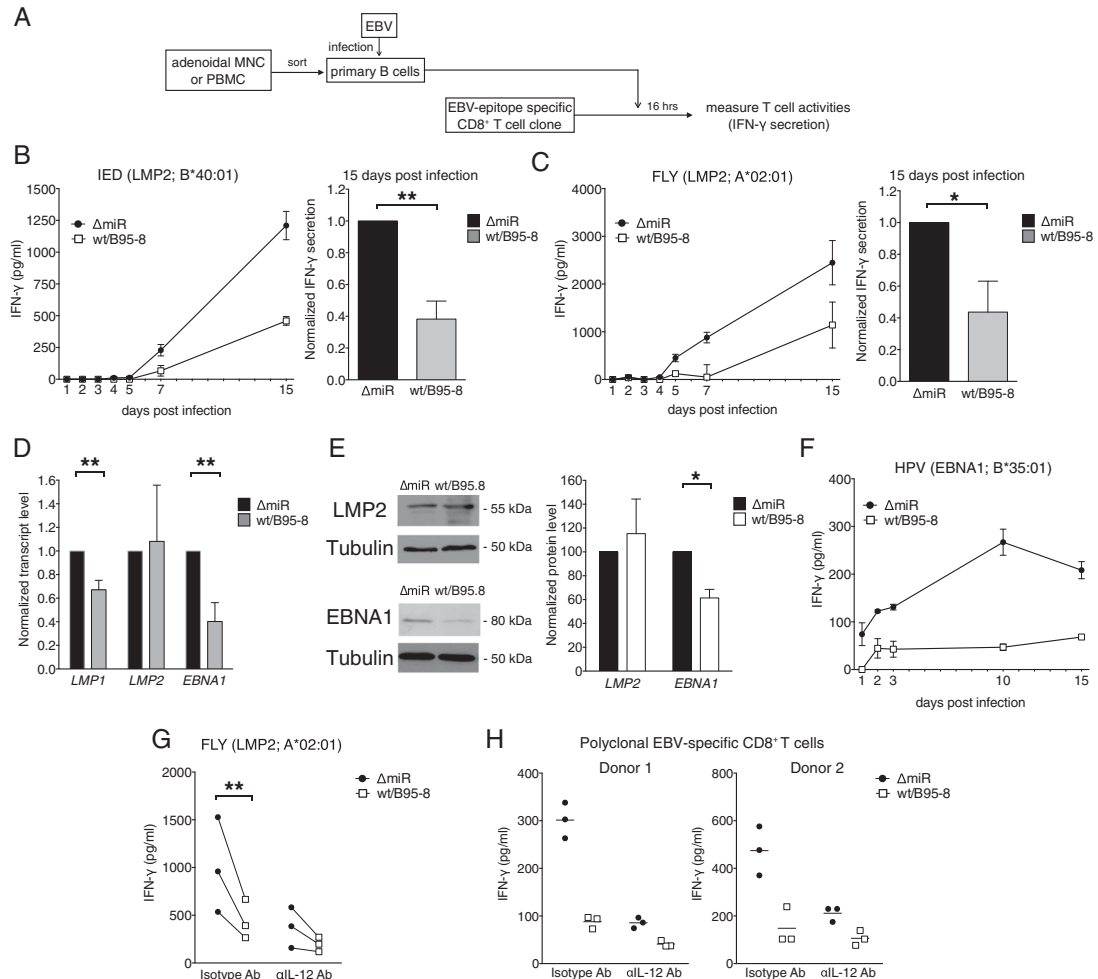


Fig. 4. EBV miRNAs control recognition of diverse types of CD8⁺ T-cell epitopes. (A) Schematic overview of the experiments shown in B, C, F, and G of this figure. The presentation of viral epitopes from B cells infected with WT/B95-8 or Δ miR EBV was analyzed with epitope-specific CD8⁺ T-cell clones or polyclonal lines. Equal numbers of B and T cells were cocultured for 16 h, and IFN- γ release was measured by ELISA at the indicated time points. (B and C) The presentation of two LMP2 epitopes by infected B cells was analyzed with CD8⁺ T-cell clones specific for the HLA-B*40:01-restricted IED epitope (B) or the HLA-A*02:01-restricted FLY (C) epitope. A representative time course experiment with mean values and SD of three replicates (Left in B and C) and the summary of all experiments performed 15 dpi with B cells from five different donors (Right in B and C) are shown. Results are normalized to values of Δ miR EBV-infected cells. (D) Relative transcript levels of LMP1, LMP2, and EBNA1 were assessed by quantitative RT-PCR in B cells infected with WT/B95-8 or Δ miR EBV at day 15 post infection. Transcript levels are normalized as in Fig. 3A. LMP1 is a known target of viral miRNAs (43, 44) and was used here as a positive control. Data are shown as mean values and SD of five donors. (E) Western blot analysis of EBNA1 and LMP2 in B cells infected with WT/B95-8 or Δ miR EBV 15 dpi. Representative examples (Left) and protein levels relative to Tubulin (Right) are shown. The result from Δ miR EBV-infected cells was set to 100%. Data are shown as mean values and SD of four different donors. (F) Recognition of the HLA-B*35:01-restricted EBNA1 epitope HPV presented by infected B cells as in B and C. (G) IL-12 neutralization reduces the activation of an epitope-specific CD8⁺ T-cell clone. Infected B cells (15 dpi) were cocultured as in C (Right) with the FLY-specific T-cell clone together with an anti-IL12B antibody or a control antibody of the same isotype (2.5 μ g/ml). After 16 h, IFN- γ release was measured by ELISA. An overview of three experiments with different donors is shown. (H) Equal numbers of infected B cells and polyclonal EBV-specific CD8⁺ T cells were cocultured together with an anti-IL12B antibody or a control antibody of the same isotype (2.5 μ g/ml). After 16 h, IFN- γ release was measured by ELISA. Results from two different donors are shown. * P < 0.05. ** P < 0.01.

was not affected by the expression of EBV miRNAs (Fig. 3F). These experiments strongly supported the presence of additional immunoevasive mechanisms that affect recognition of the FLY epitope. To address the possibility that LMP2A/B gene expression may be regulated by EBV miRNAs, we evaluated the gene transcript levels

by quantitative RT-PCR in infected B cells 15 d post infection and found LMP2A/B unaffected by viral miRNAs (Fig. 4D). In contrast, LMP1, a known target of EBV miRNAs (43, 44), was down-regulated as expected (Fig. 4D). Similarly, LMP2 protein levels did not depend on viral miRNAs (Fig. 4E), substantiating the conclusion

that EBV miRNAs regulate the activation of LMP2A/B epitope-specific T-cell clones without affecting the viral source of the epitopes they recognize.

In contrast to LMP2A, EBNA1 transcripts appeared to be under the control of viral miRNAs (Fig. 4D) (35), also resulting in decreased levels of EBNA1 protein 15 d post infection (Fig. 4E). We therefore tested the recognition of the HPV epitope of EBNA1 presented by the HLA-B*35:01 allele in a time course experiment (Fig. 4F). After only 1 day post infection, the HPV-specific T-cell clone was clearly activated but only when challenged with ΔmiR EBV-infected B cells. Later on, B cells infected with either virus presented the HPV peptide, but ΔmiR EBV-infected B cells were preferentially recognized (Fig. 4F). Regulation of TAP (Fig. 3A–C); lower surface levels of the presenting HLA-B allele, which might be down-regulated similar to other HLA-B allotypes (Fig. 3F); and reduced levels of EBNA1 gene expression (Fig. 4E) may all have contributed to this result.

The analysis of EBNA1 epitope presentation did not reveal why TAP-independent LMP2A-derived peptides presented via HLA-A*0201 were under the control of viral miRNAs (Fig. 4C). We speculated that other costimulatory molecules or proinflammatory cytokines (35) may be responsible for this TAP-independent immunoevasive function. In particular, IL-12, which contributes to the activation of effector T-cell functions (45), was a possible candidate because it is a direct and prominent target of at least five different EBV miRNAs 5 d post infection (35) and was also down-regulated at the transcript level 15 d post infection (Fig. 3A). To address this possibility, we neutralized IL-12 secreted from EBV-infected B cells with a suitable antibody and measured IFN-γ secretion by the FLY-specific T-cell clone (Fig. 4G). IL-12 neutralization dramatically reduced the activation of the T cells cocultured with ΔmiR EBV-infected cells. T-cell activation with WT/B95-8 EBV-infected target cells was also affected but to a lesser extent. Very similar results were observed with polyclonal EBV-specific CD8⁺ T cells cocultured with HLA-matched infected B cells (Fig. 4H), showing that miRNA-mediated regulation of IL-12 was globally decreasing recognition by CD8⁺ T cells. This effect, which was clearly evident for EBV-specific CD8⁺ T cells, was mild with EBV-specific CD4⁺ T cells (Fig. S3).

Discussion

In this study, we show that EBV miRNAs inhibit surveillance of EBV by CD8⁺ T cells. Viral miRNAs reduce virus-specific proliferation, cytokine production, and killing of infected cells by CD8⁺ T cells with various EBV latent epitope specificities. We identified several mechanisms for this inhibition. First, miRNAs target *TAP2* directly, down-regulate the entire TAP complex, and reduce HLA allotypes that preferentially present TAP-dependent epitopes. Second, miRNAs repress EBNA1, which limits the level of a protein but is essential during most forms of EBV latency. Third, miRNAs diminish IL-12 release by infected B cells, reducing the virus-specific activity of EBV-specific CD8⁺ T cells. Thus, EBV miRNAs limit surveillance by CD8⁺ T cells through multiple mechanisms, likely contributing to the maintenance of lifelong infection.

It is an attractive hypothesis (35) that T-cell immunoevasion in latency would be most economically achieved by miRNAs due to their nonantigenicity. This hypothesis is now fully substantiated by our present findings that EBV miRNAs interfere with several steps of antigen presentation preventing CD8⁺ T-cell recognition of latently infected B cells. These results are complementary to our previous findings documenting that EBV miRNAs regulate multiple pathways important for differentiation and activation of antiviral CD4⁺ T cells in the first days of infection (35). That study also provided some evidence that CD4⁺ T-cell recognition is also regulated later, as the structural protein gp350 could be detected by CD4⁺ T cells in cells 15 d post infection but only

when miRNAs were absent (35). The mechanisms of regulation we identified in that context—that is, regulation of MHC II, lysosomal enzymes, and IL-12—are likely relevant in latency as well and may explain why EBNA-specific CD4⁺ T cells are generally impaired in recognizing LCLs (46). The present study focused on established latency, but because we observed a strong miRNA regulation of EBNA1 recognition by CD8⁺ T cells already on days 1, 2, and 3 after infection, the hypothesis that EBV miRNAs generally suppress CD8⁺ T-cell recognition already in the first days of infection during prelatency (13, 22, 47) deserves closer investigation in the future.

An overview of EBV miRNAs that directly target pathways involved in CD8⁺ and/or CD4⁺ T-cell recognition of infected B cells is provided in Table 2. As large subsets of viral miRNAs are expressed in all phases of EBV's life cycle (48), it appears plausible that viral miRNAs inhibit these target molecules globally. For example, miRNAs miR-BART1, miR-BART2, and miR-BART22, which target IL-12 (35), are all highly expressed not only in latency III but also in EBV-infected germinal center B cells (latency II) and memory B cells (latency 0/I) from healthy donors as well as different types of EBV-associated cancer cells (34, 48, 49). Therefore, miRNA-mediated reduction of IL-12 could lead to decreased T-cell activation and recognition at different stages of infection and malignant disease. Among TAP-regulating miRNAs, miR-BHRF1–3 is predominantly expressed initially upon infection and in latency III in vitro (36), but miR-BART17 also shows expression in memory B cells and cancer cells (48), suggesting that EBV miRNA-mediated TAP regulation could likewise be important in vivo.

Although EBV-specific immunity is likely to operate at different stages of latency and lytic replication to control viral infection, the question of interest is whether latency III, in its own right, is a target of EBV-specific immunosurveillance by T cells. For immunosuppressed patients, it appears clear that T-cell deficiency favors appearance of latency III malignancies (3), that adoptive T-cell therapy can prevent this (4, 5), and that T-cell therapy fails if the EBV strain in question does not express crucial CD8⁺ T-cell epitopes in a latency III protein (50). Regarding infection in immunocompetent carriers, there were early arguments against a T-cell surveillance of latency III (51), but later studies showed that latency III-associated CD8⁺ T-cell epitopes are in fact under a selective pressure that depends on the frequency of HLA class I allotypes in a population (52, 53). Cumulatively, these reports suggest that EBV-specific CD8⁺ T-cell surveillance of latency III is an important aspect of infection control in vivo.

In this work, we have analyzed expansion of EBV-specific T cells, cytokine secretion, and cytolysis to study the interference of EBV miRNAs with CD8⁺ T-cell functions. We found such

Table 2. Direct targets of EBV miRNAs with immune functions

Function	Gene (protein)	EBV miRNAs
Antigen processing	<i>CTSB</i> (Cathepsin B)	BHRF1–2 BART2–5p BART2–5p
	<i>LGGMN</i> (AEP)	BART1–5p
	<i>IFI30</i> (GILT)	BART1–3p BART1–3p
Peptide transport	<i>TAP2</i> (TAP)	BHRF1–3 BART17
Cytokines	<i>IL12B</i> (IL-12p40)*	BHRF1–2 BART1–3p BART2–5p BART10–3p BART22

*Component of IL-12 (p35/p40) and IL-23 (p19/p40).

interference for T cells specific for five out of five epitopes from three different antigens (LMP2A/B, EBNA1, and EBNA3A). Because this collection contained epitopes with different HLA class I restrictions, derived from different categories of antigen (nuclear vs. transmembrane proteins) and with different processing requirements (TAP-dependent or -independent, proteasome- or immunoproteasome-dependent), our data indicate that the cumulative functional impact of miRNAs allows latently infected B cells to hide from CD8⁺ T cells in general. Our findings with polyclonal CD8⁺ T cells of complex composition corroborated our view.

miRNAs mediate their effects through direct binding to their target transcripts such as *TAP2*, for example (Fig. 3C). Because the recognition of TAP-independent epitopes is also inhibited by miRNAs, other mechanisms must contribute to the reduced presentation of the TAP-independent LMP2 epitopes CLG (Fig. 2G) and FLY (Fig. 4). Only a minority of all CD8⁺ T-cell epitopes (including those of viral origin) are expected to be TAP-independent (42), and the importance of TAP is reflected by the many herpesviruses that have evolved their own TAP-inhibitory proteins (54, 55). A broad impact of TAP regulation on the immunological status in latency is also suggested by our observation that viral miRNAs do not affect global levels of HLA-A2, which is capable of presenting highly hydrophobic peptides that are more likely to be TAP-independent (39, 56), but do reduce levels of the HLA class I allotypes tested (HLA-B7, B8, and B40). These and most HLA class I allotypes are less likely to present TAP-independent epitopes, because they require the presence of polar or charged anchor residues in the peptide (57, 58).

Another effect of EBV miRNAs, suppression of IL-12, seems to act globally on the function of antigen-specific T cells. Although IL-12 was originally identified as a product of EBV-infected LCLs (59), its role in EBV-specific CD8⁺ T-cell immunity has remained obscure. In addition to its well-known function in promoting Th1 differentiation (60), IL-12 was shown in mouse and human studies to promote CD8⁺ T-cell functions such as proliferation, cytolysis, and IFN- γ production (60, 61) through STAT4 signaling, up-regulating T-bet, and increasing IL-2 sensitivity (62, 63). In our experiments, blockade of IL-12 fully reverted the effect of the miRNA deletion for polyclonal EBV-specific CD8⁺ T cells (Fig. 4G and H) but unexpectedly had only a minor effect on polyclonal EBV-specific CD4⁺ T cells (Fig. S3). The reason for this difference is not clear yet, but one possibility may be a differential requirement for costimulatory signals (64). However, as we have shown (35), EBV miRNAs affect CD4⁺ T-cell responses through IL-12 regulation already at the level of differentiation from naive T cells and thus act on both major classes of T cells. Because we found IL-12 to be the gene product most strongly down-regulated by EBV miRNAs, control of this cytokine appears to be central to the maintenance of EBV infection.

Materials and Methods

Patient Samples. PBMCs and surgically removed adenoid biopsies were obtained from volunteer blood donors and patients from the Department of Otorhinolaryngology of the Universitätsklinikum der Ludwig-Maximilians-Universität München, respectively. The local ethics committee (Ethikkommission bei der Ludwig-Maximilians-Universität München) approved the use of this human material. Informed consent was not required because the biopsies originated from disposed tissues from anonymous donors who underwent routine surgery.

Human Primary Cells, Cell Lines, and Cell Culture. Human primary B and T cells were prepared from adenoidal mononuclear cells (MNCs) or PBMCs as described (35). The EBV-positive Burkitt's lymphoma cell line Raji, HEK293-based EBV producer cell lines, infected human primary B cells, LCLs, and isolated T cells were cultivated as described in *SI Materials and Methods*.

Preparation of EBV Stocks and Infection of Human Primary B Cells. Stocks of recombinant EBV strains were essentially prepared and quantitated as described (65). Details can be found in *SI Materials and Methods*.

In Vitro Model of EBV Infection (B-Cell Outgrowth Assay). B cells (CD19⁺) were isolated from PBMCs of EBV-positive donors and infected with WT/B95-8 or Δ miR EBV strains. After 12 h, the infected B cells were extensively washed to remove free virions. CD8⁺ T cells isolated from the same donors were cocultured at B:T cell ratios ranging from 8:1–1:2 (seeding 32,000 B cells per well). B cells cultivated without T cells served as control. Cells were refed weekly. After 4 wk, the cultures were analyzed for viable cells in MTT assays as previously described (22).

Establishment of EBV-Specific Effector T Cells and T-Cell Clones. EBV-specific CD8⁺ T-cell clones were established by limiting dilution from polyclonal T-cell lines that were generated by stimulating PBMCs with LCLs infected with WT/B95-8 or from specific T cells directly obtained from peripheral blood cells by peptide stimulation, IFN- γ capture (Miltenyi Biotec), and magnetic isolation (66, 67). Likewise, EBV-specific CD4⁺ T cells were generated by repetitive stimulation of sorted CD4⁺ T cells with autologous LCLs infected with WT/B95-8 EBV as described previously (66).

EBV-Specific T-Cell Recognition. EBV-specific effector T-cell activities were measured with IFN- γ ELISA and calcein release assays. For IFN- γ detection from T cells, effector and target cells were cocultured at a 1:1 ratio (5×10^4 cell per well) for 16 h in a 96-well plate (V bottom). IFN- γ levels were detected with ELISA following the manufacturer's protocol (Mabtech). IFN- γ concentrations below 16 pg/mL were regarded negative. Neutralization of IL-12 was performed with an antibody (2.5 μ g/mL), which was added directly to the coculture and is directed against the p40 subunit of IL-12 (C8.6; BioLegend). An analogous isotype control antibody (MOPC-21; BioLegend) was used as a control.

T-Cell Cytotoxicity Assays. EBV-infected B cells were purified by Ficol-Hypaque (PAN-Biotech) gradient centrifugation, and 5×10^5 target cells were labeled with calcein (Invitrogen) at 0.5 μ g/mL. After three washing steps with PBS, target and effector cells were cocultured in V bottom 96-well plates with different ratios in RPMI without Phenol Red (PAN-Biotech). After 4 h of coculture, fluorescence intensities in supernatants were measured by the Infinite F200 PRO fluorometer (Tecan). As controls, spontaneous calcein release of target cells cultivated without effector cells and cells lysed with 0.5% Triton-X100 (Carl Roth) were used to define the levels of no and fully lysed target cells, respectively.

T-Cell Expansion Assay. CD8⁺ T cells were isolated from PBMCs of EBV-positive donors. We stimulated 1×10^6 CD8⁺ T cells with 1×10^5 autologous irradiated B cells (infected for 15 d) and 20 U/mL IL-2. Cells were restimulated every 10 d. At 10 and 20 d after the first stimulation, T cells were stained with unlabeled HLA/peptide pentamers (Proimmune) for 20 min at 37 °C. Counterstaining was done with CD8 and CD3-specific antibodies and Pro5 fluorotag (Proimmune) on ice for 30 min. T-cell numbers were determined using calibrated APC-beads as volume standard by flow cytometry (68).

Luciferase Reporter Assays. Details of the reporter plasmids and the technical aspects of the dual luciferase reporter assays can be found in *SI Materials and Methods*.

Quantitative RT-PCR. Isolation of RNAs and their analyses by PCR are described in *SI Materials and Methods*.

Western Blotting. Cell lysis and antibodies used to detect viral and cellular proteins of interest can be found in *SI Materials and Methods*.

Flow Cytometry and Antibodies. Techniques and antibodies used to detect various surface molecules are described in detail in *SI Materials and Methods*.

Statistical Analysis. We used Prism 6.0 software (GraphPad) for statistical analysis, and the two-tailed ratio *t* test was applied unless otherwise mentioned.

ACKNOWLEDGMENTS. We thank Elisabeth Kremmer and Dagmar Pich for monoclonal antibodies and valuable experimental advice, respectively. This work was financially supported by grants of the Deutsche Forschungsgemeinschaft (SFB1054/TP B05, SFB1064/TP A13, SFB-TR36/TP A04), Deutsche Krebshilfe (107277 and 109661), National Cancer Institute (CA70723 and CA022443), and personal grants to T.T. from Deutscher Akademischer Austauschdienst (DAAD, Studienstipendien für ausländische Graduierte aller wissenschaftlichen Fächer) and to M.B. from the European Molecular Biology Organization (EMBO).

9.2. EFFECT OF EPSTEIN-BARR VIRAL MIRNAS ON CD4⁺ AND CD8⁺ T CELLS

1. van Baarle D, et al. (2001) Dysfunctional Epstein-Barr virus (EBV)-specific CD8(+) T lymphocytes and increased EBV load in HIV-1 infected individuals progressing to AIDS-related non-Hodgkin lymphoma. *Blood* 98(1):146–155.
2. Smets F, et al. (2002) Ratio between Epstein-Barr viral load and anti-Epstein-Barr virus specific T-cell response as a predictive marker of posttransplant lymphoproliferative disease. *Transplantation* 73(10):1603–1610.
3. Landgren O, et al. (2009) Risk factors for lymphoproliferative disorders after allogeneic hematopoietic cell transplantation. *Blood* 113(20):4992–5001.
4. Heslop HE, et al. (2010) Long-term outcome of EBV-specific T-cell infusions to prevent or treat EBV-related lymphoproliferative disease in transplant recipients. *Blood* 115(5):925–935.
5. Moosmann A, et al. (2010) Effective and long-term control of EBV PTLD after transfer of peptide-selected T cells. *Blood* 115(14):2960–2970.
6. Hislop AD, Taylor GS, Saucedo D, Rickinson AB (2007) Cellular responses to viral infection in humans: Lessons from Epstein-Barr virus. *Annu Rev Immunol* 25:587–617.
7. Bihl F, et al. (2006) Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses. *J Immunol* 176(7):4094–4101.
8. Kalla M, Göbel C, Hammerschmidt W (2012) The lytic phase of Epstein-Barr virus requires a viral genome with 5-methylcytosine residues in CpG sites. *J Virol* 86(1):447–458.
9. Thorley-Lawson DA, Gross A (2004) Persistence of the Epstein-Barr virus and the origins of associated lymphomas. *N Engl J Med* 350(13):1328–1337.
10. Rowe M, Lear AL, Croom-Carter D, Davies AH, Rickinson AB (1992) Three pathways of Epstein-Barr virus gene activation from EBNA1-positive latency in B lymphocytes. *J Virol* 66(1):122–131.
11. Longnecker RM, Kieff E, Cohen JI (2013) In *Fields Virology*, eds Knipe DM, et al. (Wolters Kluwer, Philadelphia), Vol 2, pp 1898–1959.
12. Joseph AM, Babcock GJ, Thorley-Lawson DA (2000) Cells expressing the Epstein-Barr virus growth program are present in and restricted to the naive B-cell subset of healthy tonsils. *J Virol* 74(21):9964–9971.
13. Hammerschmidt W (2015) The epigenetic life cycle of Epstein-Barr virus. *Curr Top Microbiol Immunol* 390(Pt 1):103–117.
14. Hislop AD, et al. (2007) A CD8+ T cell immune evasion protein specific to Epstein-Barr virus and its close relatives in Old World primates. *J Exp Med* 204(8):1863–1873.
15. Horst D, et al. (2009) Specific targeting of the EBV lytic phase protein BNLF2a to the transporter associated with antigen processing results in impairment of HLA class I-restricted antigen presentation. *J Immunol* 182(4):2313–2324.
16. Zuo J, et al. (2009) The Epstein-Barr virus G-protein-coupled receptor contributes to immune evasion by targeting MHC class I molecules for degradation. *PLoS Pathog* 5(1):e1000255.
17. Zuo J, et al. (2011) The Epstein-Barr virus-encoded BILF1 protein modulates immune recognition of endogenously processed antigen by targeting major histocompatibility complex class I molecules trafficking on both the exocytic and endocytic pathways. *J Virol* 85(4):1604–1614.
18. Rowe M, et al. (2007) Host shutoff during productive Epstein-Barr virus infection is mediated by BGLF5 and may contribute to immune evasion. *Proc Natl Acad Sci USA* 104(9):3366–3371.
19. Zuo J, et al. (2008) The DNase of gammaherpesviruses impairs recognition by virus-specific CD8+ T cells through an additional host shutoff function. *J Virol* 82(5):2385–2393.
20. Quinn LL, et al. (2015) The missing link in Epstein-Barr Virus immune evasion: The BDLF3 gene induces ubiquitination and downregulation of major histocompatibility complex class I (MHC-I) and MHC-II. *J Virol* 90(1):356–367.
21. Croft NP, et al. (2009) Stage-specific inhibition of MHC class I presentation by the Epstein-Barr virus BNLF2a protein during virus lytic cycle. *PLoS Pathog* 5(6):e1000490.
22. Jochum S, Moosmann A, Lang S, Hammerschmidt W, Zeidler R (2012) The EBV immunoevasin vIL-10 and BNLF2a protect newly infected B cells from immune recognition and elimination. *PLoS Pathog* 8(5):e1002704.
23. Cassell DJ, Schwartz RH (1994) A quantitative analysis of antigen-presenting cell function: Activated B cells stimulate naive CD4 T cells but are inferior to dendritic cells in providing costimulation. *J Exp Med* 180(5):1829–1840.
24. Wiesner M, et al. (2008) Conditional immortalization of human B cells by CD40 ligation. *PLoS One* 3(1):e1464.
25. Rowe M, et al. (1991) Epstein-Barr virus (EBV)-associated lymphoproliferative disease in the SCID mouse model: Implications for the pathogenesis of EBV-positive lymphomas in man. *J Exp Med* 173(1):147–158.
26. Lautscham G, et al. (2003) Identification of a TAP-independent, immunoproteasome-dependent CD8+ T-cell epitope in Epstein-Barr virus latent membrane protein 2. *J Virol* 77(4):2757–2761.
27. Hill AB, et al. (1995) Class I major histocompatibility complex-restricted cytotoxic T lymphocytes specific for Epstein-Barr virus (EBV)-transformed B lymphoblastoid cell lines against which they were raised. *J Exp Med* 181(6):2221–2228.
28. Pfeffer S, et al. (2004) Identification of virus-encoded microRNAs. *Science* 304(5671):734–736.
29. Pfeffer S, et al. (2005) Identification of microRNAs of the herpesvirus family. *Nat Methods* 2(4):269–276.
30. Cai X, et al. (2006) Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathog* 2(3):e23.
31. Kincaid RP, Sullivan CS (2012) Virus-encoded microRNAs: An overview and a look to the future. *PLoS Pathog* 8(12):e1003018.
32. Haneklaus M, et al. (2012) Cutting edge: MIR-223 and EBV miR-BART15 regulate the NLRP3 inflammasome and IL-1 β production. *J Immunol* 189(8):3795–3799.
33. Nachmani D, Stern-Ginossar N, Sarid R, Mandelboim O (2009) Diverse herpesvirus microRNAs target the stress-induced immune ligand MICB to escape recognition by natural killer cells. *Cell Host Microbe* 5(4):376–385.
34. Xia T, et al. (2008) EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-miR-BHRF1-3. *Cancer Res* 68(5):1436–1442.
35. Tagawa T, et al. (2016) Epstein-Barr viral miRNAs inhibit antiviral CD4+ T-cell responses targeting IL-12 and peptide processing. *J Exp Med*. 10.1084/jem.20160248.
36. Seto E, et al. (2010) Micro RNAs of Epstein-Barr virus promote cell cycle progression and prevent apoptosis of primary human B cells. *PLoS Pathog* 6(8):e1001063.
37. Feederle R, et al. (2011) A viral microRNA cluster strongly potentiates the transforming properties of a human herpesvirus. *PLoS Pathog* 7(2):e1001294.
38. Feederle R, et al. (2011) The members of an Epstein-Barr virus microRNA cluster cooperate to transform B lymphocytes. *J Virol* 85(19):9801–9810.
39. Lautscham G, et al. (2001) Processing of a multiple membrane spanning Epstein-Barr virus protein for CD8(+) T cell recognition reveals a proteasome-dependent, transporter associated with antigen processing-independent pathway. *J Exp Med* 194(8):1053–1068.
40. Dölken L, et al. (2010) Systematic analysis of viral and cellular microRNA targets in cells latently infected with human gamma-herpesviruses by RISC immunoprecipitation assay. *Cell Host Microbe* 7(4):324–334.
41. Garcia DM, et al. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat Struct Mol Biol* 18(10):1139–1146.
42. Del Val M, Lázaro S, Ramos M, Antón LC (2013) Are membrane proteins favored over cytosolic proteins in TAP-independent processing pathways? *Mol Immunol* 55(2):117–119.
43. Lo AK, et al. (2007) Modulation of LMP1 protein expression by EBV-encoded microRNAs. *Proc Natl Acad Sci USA* 104(41):16164–16169.
44. Riley KJ, et al. (2012) EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO J* 31(9):2207–2221.
45. Gately MK, Wolitzky AG, Quinn PM, Chizzonite R (1992) Regulation of human cytolytic lymphocyte responses by interleukin-12. *Cell Immunol* 143(1):127–142.
46. Long HM, et al. (2005) CD4+ T-cell responses to Epstein-Barr virus (EBV) latent-cycle antigens and the recognition of EBV-transformed lymphoblastoid cell lines. *J Virol* 79(8):4896–4907.
47. Jochum S, Ruis R, Moosmann A, Hammerschmidt W, Zeidler R (2012) RNAs in Epstein-Barr virus control early steps of infection. *Proc Natl Acad Sci USA* 109(21):E1396–E1404.
48. Qiu J, et al. (2011) A novel persistence associated EBV miRNA expression profile is disrupted in neoplasia. *PLoS Pathog* 7(8):e1002193.
49. Kim DN, et al. (2007) Expression of viral microRNAs in Epstein-Barr virus-associated gastric carcinoma. *J Virol* 81(2):1033–1036.
50. Gottschalk S, et al. (2001) An Epstein-Barr virus deletion mutant associated with fatal lymphoproliferative disease unresponsive to therapy with virus-specific CTLs. *Blood* 97(4):835–843.
51. Khanna R, et al. (1997) Evolutionary dynamics of genetic variation in Epstein-Barr virus isolates of diverse geographical origins: Evidence for immune pressure-independent genetic drift. *J Virol* 71(11):8340–8346.
52. Midgley RS, Bell AI, McGeoch DJ, Rickinson AB (2003) Latent gene sequencing reveals familial relationships among Chinese Epstein-Barr virus strains and evidence for positive selection of A11 epitope changes. *J Virol* 77(21):11517–11530.
53. Midgley RS, et al. (2003) HLA-A11-restricted epitope polymorphism among Epstein-Barr virus strains in the highly HLA-A11-positive Chinese population: Incidence and immunogenicity of variant epitope sequences. *J Virol* 77(21):11507–11516.
54. Rensing ME, Luteijn RD, Horst D, Wiertz EJ (2013) Viral interference with antigen presentation: Trapping TAP. *Mol Immunol* 55(2):139–142.
55. Verweij MC, et al. (2015) Viral inhibition of the transporter associated with antigen processing (TAP): A striking example of functional convergent evolution. *PLoS Pathog* 11(4):e1004743.
56. Weinzierl AO, et al. (2008) Features of TAP-independent MHC class I ligands revealed by quantitative mass spectrometry. *Eur J Immunol* 38(6):1503–1510.
57. Sutton J, et al. (1993) A sequence pattern for peptides presented to cytotoxic T lymphocytes by HLA B8 revealed by analysis of epitopes and eluted peptides. *Eur J Immunol* 23(2):447–453.
58. Falk K, et al. (1995) Peptide motifs of HLA-B58, B60, B61, and B62 molecules. *Immunogenetics* 41(2-3):165–168.
59. Kobayashi M, et al. (1989) Identification and purification of natural killer cell stimulatory factor (NKSF), a cytokine with multiple biologic effects on human lymphocytes. *J Exp Med* 170(3):827–845.
60. Manetti R, et al. (1994) Interleukin 12 induces stable priming for interferon gamma (IFN-gamma) production during differentiation of human T helper (Th) cells and transient IFN-gamma production in established Th2 cell clones. *J Exp Med* 179(4):1273–1283.
61. Valiante NM, Rengaraju M, Trinchieri G (1992) Role of the production of natural killer cell stimulatory factor (NKSF/IL-12) in the ability of B cell lines to stimulate T and NK cell proliferation. *Cell Immunol* 145(1):187–198.
62. Kaech SM, Cui W (2012) Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* 12(11):749–761.
63. Starbeck-Miller GR, Xue HH, Harty JT (2014) IL-12 and type I interferon prolong the division of activated CD8 T cells by maintaining high-affinity IL-2 signaling in vivo. *J Exp Med* 211(1):105–120.
64. Elloso MM, Scott P (2001) Differential requirement of CD28 for IL-12 receptor expression and function in CD4(+) and CD8(+) T cells. *Eur J Immunol* 31(2):384–395.
65. Steinbrück L, et al. (2015) K1 and K15 of Kaposi's sarcoma-associated herpesvirus are partial functional homologues of latent membrane protein 2A of Epstein-Barr virus. *J Virol* 89(14):7248–7261.
66. Adhikary D, et al. (2007) Immunodominance of lytic cycle antigens in Epstein-Barr virus-specific CD4+ T cell preparations for therapy. *PLoS One* 2(7):e583.
67. Rancan C, Schirmann L, Hüls C, Zeidler R, Moosmann A (2015) Latent membrane protein LMP2A impairs recognition of EBV-infected cells by CD8+ T cells. *PLoS Pathog* 11(6):e1004906.
68. Iskra S, Kalla M, Delecluse HJ, Hammerschmidt W, Moosmann A (2010) Toll-like receptor agonists synergistically increase proliferation and activation of B cells by Epstein-Barr virus. *J Virol* 84(7):3612–3623.
69. Delecluse HJ, Hilsendegen T, Pich D, Zeidler R, Hammerschmidt W (1998) Propagation and recovery of intact, infectious Epstein-Barr virus from prokaryotic to human cells. *Proc Natl Acad Sci USA* 95(14):8245–8250.
70. Haraguchi T, Ozaki Y, Iba H (2009) Vectors expressing efficient RNA decoys achieve the long-term suppression of specific microRNA activity in mammalian cells. *Nucleic Acids Res* 37(6):e43.

9.3 DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences



Until today, the accurate analysis of small RNA populations remains an unsolved problem. Due to the special characteristics of these short RNA sequences major obstacles arise when these sequences map to multiple locations in the genome, align to regions that are not annotated or underwent post-transcriptional changes. In this project, that was conducted in collaboration with several related institutes, we designed and implemented an R package for small non-coding RNA (sncRNA) profiling and applied it on 150 human and mouse samples.



The underlying analysis strategy is based on differential expression of unique sequences without prior mapping of reads to a reference genome. DEUS summarizes reads according to their actual nucleotide sequence which allows a concise analysis of the generated count matrix without ignoring reads that cannot be mapped to known feature locations. After differential expression analysis, all sequences are annotated against user defined feature databases. Finally, the complete set of sequences are clustered by similarity which allows feature-based interpretation similar to other tools but additionally provides unique insight into potential processing and editing steps among members of the same cluster.

This approach circumvents problems that typically arise when using commonly-used mapping-based approaches. Using the available sequencing data, we could show that these issues can ultimately lead to data loss and inaccurate results. The validation on 150 sequencing samples additionally revealed that multi-mapping is predominant in tissues with a potential carrier function, such as plasma, sperm and exosomes when compared to somatic cell types and colorectal cancer cell lines.

In summary, DEUS provides an unprecedented way to profile and visualize sncRNA data and considerably improves the analysis of sncRNA-seq data, being applicable in various existing pipelines and returning intuitively interpretable results.

Gene expression

DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences

Tim Jeske ^{1,2,*}, Peter Huypens^{3,4,†}, Laura Stirn^{4,5}, Selina Höckele^{3,4}, Christine M. Wurmser⁶, Anja Böhm^{4,5}, Cora Weigert^{4,5}, Harald Staiger^{3,4,5}, Christoph Klein², Johannes Beckers^{3,4,7} and Maximilian Hastreiter ^{8,9,*}

¹Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg 85764, Germany,

²Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, München 80337, Germany, ³Institute of Experimental Genetics, Helmholtz Zentrum München GmbH, Neuherberg 85764, Germany,

⁴German Center for Diabetes Research (DZD), Neuherberg 85764, Germany, ⁵Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Zentrum München at the University of Tübingen, Tübingen 72076, Germany,

⁶Chair of Animal Breeding, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising 85354, Germany, ⁷Chair of Experimental Genetics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising 85354, Germany, ⁸Institute of Computational Biology, Helmholtz Zentrum München GmbH, Neuherberg 85764, Germany and ⁹Chair of Genome-oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising 85354, Germany

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Bonnie Berger

Received on January 11, 2019; revised on June 5, 2019; editorial decision on June 9, 2019; accepted on June 18, 2019

Abstract

Summary: Despite their fundamental role in various biological processes, the analysis of small RNA sequencing data remains a challenging task. Major obstacles arise when short RNA sequences map to multiple locations in the genome, align to regions that are not annotated or underwent post-transcriptional changes which hamper accurate mapping. In order to tackle these issues, we present a novel profiling strategy that circumvents the need for read mapping to a reference genome by utilizing the actual read sequences to determine expression intensities. After differential expression analysis of individual sequence counts, significant sequences are annotated against user defined feature databases and clustered by sequence similarity. This strategy enables a more comprehensive and concise representation of small RNA populations without any data loss or data distortion.

Availability and implementation: Code and documentation of our R package at <http://ibis.helmholtz-muenchen.de/deus/>.

Contact: tim.jeske@helmholtz-muenchen.de or hastreiter@helmholtz-muenchen.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

A general approach to analyze small non-coding RNAs (sncRNA) data encompasses the evaluation of differential expression between conditions of interest. For this purpose, several software packages, such as miRDeep (Friedländer *et al.*, 2008), tDRmapper (Selitsky and Sethupathy, 2015), sRNAlyzer (Wu *et al.*, 2017) and sRNAtoolbox (Rueda *et al.*, 2015), have been developed. A common step shared by these sncRNA profiling tools is the alignment of reads to a reference genome, followed by their annotation, feature count quantification and the subsequent statistical evaluation between experimental conditions (Anders *et al.*, 2013). However, the analysis of the expressed sncRNA populations poses several hurdles because short reads are more likely to map to multiple locations in the genome, or map to genomic coordinates that are not annotated and may deviate from the originating feature sequence due to editing and post-transcriptional processing steps. Here, we present our method that analyzes differential expression of unique sequences (DEUS) for profiling sncRNA sequence data without relying on read mapping.

2 Implementation

Our pipeline starts with the identification of unique reads in each of the input FASTQ files to generate a typical RNA-seq count matrix, but utilizing the actual read sequence instead of the gene feature as identifier. The count table is then used as input for DESeq2 (Love *et al.*, 2014) analysis to calculate statistically significant read count differences between samples from different experimental conditions. Adjusted p -values for the differentially expressed (DE) unique sequences are calculated according to the Independent Hypothesis Weighting method (Ignatiadis *et al.*, 2016) using the means of normalized counts as covariate. DE unique sequences are subsequently annotated by BLASTn (Altschul *et al.*, 1990) searches against user defined BLAST databases. Subsequently, the CD-Hit clustering algorithm (Fu *et al.*, 2012; Li and Godzik, 2006) is applied to classify significant DE reads into subgroups of similar sequences based on the percentage of sequence identity and the length of the overlapping sub-sequences as defined by the user. This additional information can be used to inspect significant sequences in groups that indicate similar biological origin. Finally, a comprehensive summary table is generated by combining results from differential expression analysis, BLASTn annotation and cluster assignment (Supplementary Table S1). To easily explore the content of the table the user can define an individual set of terms that represent feature classes of interest. The given terms will be integrated as columns each containing the number of BLAST hits that match the corresponding term. DEUS also automatically generates plots to visualize the expression intensities versus fold changes of the identified sequences and the distance of the expression profiles of the samples in analysis. Additionally, we implemented an extended approach that performs clustering and summarizes sequence counts prior to differential expression analysis to provide further insights on a more general level (see Supplementary Material). We implemented each of the described steps as customizable functions in the R package DEUS. This modular design allows the user to customize our pipeline, tailored to the specific needs of the project.

3 Discussion

In accordance with Johnson *et al.* (2016), we found that sncRNA datasets from various mouse and human biomaterials are plagued

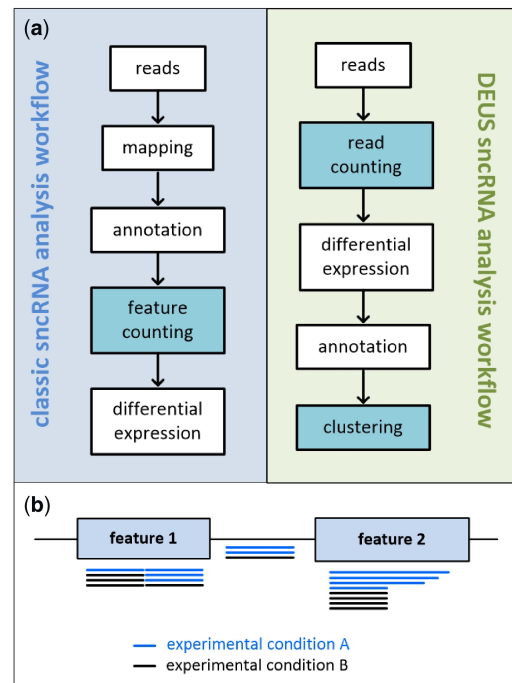


Fig. 1. Major differences between mapping-based and DEUS small RNA profiling strategies. (a) Schematic representation of the workflow of mapping-based pipelines compared with DEUS. (b) Schematic representation of scenarios that result in data distortion or data loss when applying mapping-based sncRNA profiling strategies. Mapping-based workflows ignore reads that map to non-annotated genome regions (depicted as reads between the two features) and foster data distortion as variant-specific read counts are usually summed up during subsequent feature counting even if these reads align at different spatial coordinates of the same genomic feature (depicted as reads mapped to feature 1) or exhibit discrete variations in nucleotide sequence or sequence length (depicted as reads mapped to feature 2)

by substantial amounts of multi-mapping reads ($61.5 \pm 20.1\%$) and noticeable amounts ($44.7 \pm 17.2\%$) of reads that map to regions of the genome that are not annotated (Supplementary Fig. S1 and Supplementary Table S2). Consequently, it requires dedicated methods that account for these issues. DEUS deviates from mapping-based small RNA profiling methods in several aspects (Fig. 1a). As DEUS is not relying on mapping it facilitates sncRNA profiling even when a reference genome is not available. Further, it includes all reads in analysis even those that were mapped to loci lacking feature annotation and those that cannot be mapped, for example, due to extensive RNA editing events (Fig. 1b). DEUS circumvents the challenge of correctly assigning multi-mapping reads to their originating feature by representing multiple putative mapping positions by multiple annotations per unique sequence. Due to the use of unique sequences, DEUS inherently detects discrete sequence or length variations. The information about sequence variations would otherwise be hidden in read counts grouped on feature-level or lost if varying reads could not be mapped. To allow feature-based result interpretation despite sequence-based data analysis, DEUS clusters highly similar sequences (Fig. 1a). This compression of resulting DE sequences to sequence clusters reduces the number of result entities

9.3. DEUS: AN R PACKAGE FOR ACCURATE SMALL RNA PROFILING BASED ON DIFFERENTIAL EXPRESSION OF UNIQUE SEQUENCES

DEUS

3

in a range from about 40% up to 80%. In combination with the extended differential expression analysis, the use of sequence clusters improves the overall signal detection power and provides a second data perspective that includes single sequence and cluster-level analysis.

In summary, DEUS provides an unprecedented way to profile and visualize sncRNA data. DEUS clearly diverges from mapping-based analysis strategies, hampered by substantial data loss and distortion of feature counts. We believe that our DEUS pipeline considerably improves the analysis of sncRNA-seq data, being applicable in various existing pipelines and returning intuitively interpretable results.

Funding

The work was supported by funding of The Leona M. and Harra B Helmsley Charitable Trust, the Care-for-Rare Foundation, BMBF (PID-NET, 01GM1517A) and by grants of the DZD - German Center for Diabetes Research and the Helmholtz Alliance AMPPro to JB.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.

- Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, 8, 1765–1786.
- Friedländer,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, 26, 407–415.
- Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152.
- Ignatiadis,N. *et al.* (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, 13, 577–580.
- Johnson,N.R. *et al.* (2016) Improved placement of multi-mapping small RNAs. *G3 (Bethesda)*, 6, 2103–2111.
- Li,W. and Godzik,A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659.
- Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15, 550.
- Rueda,A. *et al.* (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.*, 43, W467–W473.
- Selitsky,S.R. and Sethupathy,P. (2015) tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. *BMC Bioinformatics*, 16, 354.
- Wu,X. *et al.* (2017) sRNAAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Res.*, 45, 12140–12151.

LIST OF TABLES

1.1	Several key studies investigating schizophrenia genomics.	13
3.1	The 46 candidate <i>de novo</i> mutations found in 32 trios. For each variant, several annotations are provided. This includes the maximal allele frequency (MaxDBAltAF) across the used databases(1k201304, dbsnp138, ESP6500AA, ESP6500EA and ExAC), an indication whether the gene is located in a schizophrenia candidate (Y) locus or has an interaction with it (PPI), the CADD score as well as the affected individual.	43
3.2	Heteromeric protein interactions affected by rare interaction-disrupting mutations. For each interaction, the involved proteins, the ExAC frequency of the variant as well as the affected samples are given.	47
3.3	Candidate genes with significant schizophrenia-related mouse phenotypes ($P < 0.0001$). Along with the detected phenotype, the measurement parameter, the phenotyping center and evidence for links to schizophrenia is provided. Data obtained from the IMPC [107].	56
3.4	Long deletions detected by Pindel that affect the 22q11 region. For each deletion, the affected genes are specified including the exact region and length of the variation.	57
3.5	Excerpt of rare deleterious variants that were found to be associated with schizophrenia, either directly or by gene association. het/hom column specifies the number of cases with a heterozygous or homozygous genotype; Exon indicates the affected exon; AF is the maximal allele frequency across ExAC,1000G and ESP	57
4.1	Enriched GO terms with high redundancy. GO term descriptions according AmiGO2 [114]. The pvalue column specifies the enrichment pvalues after multiple-testing correction (WG: WebGestalt, DR: DNENRICH). Supersets indicate the GO set(s) that already covers most genes from the respective set.	74
6.1	Overview about the three different phasing strategies that were used during this work, as well as their requirements and primary phasing targets.	88

8.1 Candidate/sample overview. For each affected patient, the number of detected variants/genes is specified. CADD: Genes with increased genetic burden; LoF: Genes with *de novo* knock out; Cpht: Rare damaging compound heterozygous variants; DNM: Rare damaging *de novo* variants; EG: Genes with disrupted protein-interaction sites; CV: Variants with known schizophrenia associations based on ClinVar, SZ: schizophrenia, CSZ: Childhood-onset schizophrenia, SSZ: Susceptibility to schizophrenia; LDel: Amount of long deletion on chromosome 22; Trio(*) indicates whether parents of the patient were sequenced; Finding summarises the most important results for the respective samples. 104

LIST OF FIGURES

1.1	Spectrum of mental disorders. Various clinical syndromes are closely related by the degree of manifestation of certain symptoms. A common genetic background is shared by all syndromes. Figure taken from [20]	6
1.2	(a) Normal cortical development including proliferation, migration, arborization (circuit formation) and myelination. (b) Reduced interneuron activity and excessive excitatory pruning together with a myelination deficiency lead to an altered excitatory-inhibitory balance and reduced connectivity in the prefrontal cortex. Figure taken from [33]	9
1.3	Concordance rates for different degrees of relatedness. With decreasing relatedness, the genetic concordance rates drops from 33% -44.3% to 1%. MZ, monozygotic; DZ, dizygotic; Data taken from [4] and [41]	11
2.1	Representation of the variant processing and filtering workflow according to the applied methods. After initial quality filtering of raw data (a) , reads are processed based on GATK Best Practices (b) . The analysis-ready variants are then investigated throughout five distinct approaches (c-g) . This results in a set of candidate genes that are further analysed by system biological methods (h-i) . Subfigures b, c, e were taken from [65–67].	17
2.2	Exemplary raw quality profile showing the base quality per position index. Each bin indicates the base quality for one or multiple positions within a sequence read calculated across all available sequencing reads. Yellow boxes indicate the 25%-75% quantile with the red line indicating the median. The black whiskers show the range between the 10% and 90% quantile, the continuous blue line represents the overall mean quality.	18

2.3	Example of a Burrows-Wheeler transformation that allows efficient and reversible compression of character strings. Given a specific character combination, the BWT first generates all possible word rotations and then sorts them in lexical order. Afterwards, only the last column of the aligned character strings is stored. Due to the runs of repeated characters it can be stored efficiently and is still sufficient to reproduce the original input. The Burrows-Wheeler Aligner is based on this principle and uses it for fast and accurate short read alignment. Figure taken from [70].	19
2.4	Read mapping before and after local realignment. rs-identifiers indicate potential variant sites. Local realignment was performed for two independent data sets (1,000 Genomes Pilot 2 data mapped using MAQ [71] and HiSeq data mapped using BWA). The left panel of the plot indicates original read mapping with three possible variants (rs28782535, rs28783181 and rs28788974). After realignment around known Indels, only one of the original SNPs (rs28788974) remain. In addition, a previously hidden insertion is visible (rs34877486). In both cases, the realignment clearly improves overall mapping quality and removes false-positive variants. Figure based on [72] . . .	20
2.5	Core operations performed by HaplotypeCaller during variant calling.	21
2.6	Training of the VQSR model. A gaussian mixture model is built on training data taken from dbSNP and HapMap3. In this example, the resulting model is based on two quality scores, strand bias and variant quality score. Figure taken from [72].	22
2.7	Application of the trained VQSR model. The trained model is applied on new variant data. Based on the selected stringency, variants are filtered or kept according to their relative position compared to the known variants that were used during training. Figure taken from [72]	22
2.8	Relationship of scaled CADD scores and categorical variant consequences. A and B indicate the proportions of substitutions with a specific consequence for each scaled CADD score. C shows the score distribution of loss-of-function variants in specific gene sets: Disease genes harbor at least 5 known pathogenic mutations, Essential genes were predicted to be essential, GWAS genes harbor at least 2 loss-of-function mutations in 1000 Genomes, Olfactory genes encode the common olfactory receptor proteins and Other contains a random selection of 500 genes. Figure taken from [67]	24
2.9	Heterozygous <i>de novo</i> mutation affects offspring and is not present in either healthy parents. Figure taken from [66].	25
2.10	Compound heterozygous mutations (trans) affecting both alleles. Figure taken from [91].	26
2.11	Depending on the affected gene, Loss-of-function variants can have different phenotypic effects. While the most extreme form may lead to embryonic lethality, LoF variants may also have healthy benefits.	27
2.12	The spectrum of genetic variants ranging from common variants with low effects to rare variants with high effect sizes. Figure taken from [93]	28

2.13	Edgotypes defined by the edgetics approach. Each mutation affecting protein-protein interactions changes the interaction network in a specific manner. Depending on the outcome, different phenotypes may arise. Taken from [96]	29
2.14	The pattern growth approach as used in Pindel. While a) depicts a large deletion that happened in the sample, b) indicates a smaller insertion. Using mapping information gained from both paired reads, Pindel determines paired reads that mapped with Indels or where only one end was mapped. This information is then used to identify potential variants Figure taken from (http://gmt.genome.wustl.edu/packages/pindel/background.html).	31
2.15	User interface of the Konstanz information miner (KNIME)	33
3.1	Quality profile showing the base quality per position index for raw sequence reads, separately for sequence data from Eurofins Medigenomix GmbH and the Technical University (WZW). In both cases, raw quality scores indicate an already high overall quality. However, as expected, quality scores at the beginning as well as the end of the read drop towards or below the high quality threshold as indicated by the horizontal line (green).	37
3.2	Quality profile showing the base quality per position index for filtered sequence reads, separately for sequence data from Eurofins Medigenomix GmbH and the Technical University (WZW). In comparison to the raw quality profile, the overall quality of the remaining bases/reads has increased especially towards the read ends.	38
3.3	Beeswarm plot showing overall mapping statistics for all included samples. The left panel indicates the total number of sequence reads that were generated for each sample. The second panel presents the percentage of reads that were successfully mapped to the human reference genome (hg19). On average this was true for about 99% of the reads which equals a total amount of 81 million sequence reads. The right panel presents the percentage of reads that were mapped to primary target region (SeqCap EZ Exome v3). The two clusters that are visible in this panel can be explained by the different capture kit that was used for samples sequenced at TUM Weihenstephan.	39
3.4	Overall number of <i>de novo</i> mutations, grouped according to their genomic impact. Part (a) includes all 74 <i>de novo</i> variants, (b) the 46 rare deleterious <i>de novo</i> variants. . . .	40
3.5	Number of <i>de novo</i> variants for each affected trio offspring. Each bar represents the number of all and filtered (rare deleterious) <i>de novo</i> variants whereas the blue dotted line indicates the number of expected variants based on the neutral <i>de novo</i> mutation rate of 1.1×10^{-8} . Additionally, the age of both parents (green and red line)as well as the age of onset (black line) are shown.	41

3.6	Compound heterozygous variants in schizophrenia. (a) indicates the number of genes (coding region) affected by one or more compound heterozygous variants. Six genes were affected more than five times which, however, is most likely explained by the median protein sequence length of 6777 that is clearly above average (Median length of human proteins classified in Pfam-A: 416 [113]). (b) shows the distribution of the respective ExAC allele frequencies and CADD scores that are slightly negative correlated (-0.15 with a 95% confidence interval [-0.21,-0.10] and p-value 3.96×10^{-8}).	42
3.7	Number of rare deleterious compound heterozygous variants for trio and non-trio data. Overall, more variants were detected in trio data than in non-trio samples. This can be explained by the reduced phasing information that can be generated for individual samples.	44
3.8	Number of LoF variants for the complete and frequency-filtered variant set for schizophrenia patients. In both cases, the overall number of LoF variants, homozygous LoF variants and mutations that affect all transcripts (at the respective site) and lead to complete knockout of the gene product are given.	45
3.9	Overview about the increase of genetic burden for the 27192 canonical genes defined by UCSC. Each bar represents the number of genes for which at least n offspring samples have an increased genetic burden.	46
3.10	Exemplary genetic burden difference (CSS_{Gene} difference) between affected offsprings and the respective parents for Hornerin (HRNR), a gene involved in calcium ion binding. a) indicates individuals with increased burden, b) with equal burden and c) with decreased burden.	46
3.11	Interaction between ACY3 (Aminoacylase 3) and ASPA (Aspartoacylase). The mutated residue is shown in green, the remaining PPI interface residues in red. Here, the mutated Proline at position 181 of the Aspartoacylase protein affects the direct interaction with Aminoacylase 3.	47
3.12	Number of candidate genes resulting from the different analysis methods (a) and the respective number of affected samples (b). As samples are partially affected by multiple variant types, there exist sample overlaps between certain affection groups. For individuals grouped into "None", I did not find any potential disease-relevant variants.	48
3.13	Excerpt of the STRING interaction network using the 468 candidate genes and direct interaction partners (limited to 99). Interaction sources were restricted to experiments and databases with at least high confidence (STRING). Candidate genes are highlighted in red; the size of the node indicates the number of interaction partners.	49

3.14	Enriched GO terms for the three GO categories (Molecular function, Cellular component, Biological process). Microtubule, actin and motor protein related terms are marked with a red *. For each category the respective GO ID and number of candidate genes falling into this category (n) are given. Hatched bar indicates that the group was discarded after DNENRICH analysis.	50
3.15	Enriched GO terms after DNENRICH analysis. Microtubule, actin and motor protein related terms are marked with a red *. For each category the number of candidate genes falling into this category (n) are given. Hatched bar indicates that the group not significantly enriched. The dashed line indicates the applied p-value cutoff.	51
3.16	Damaging genetic variants in genes of the neuronal cytoskeleton. For the 288 samples with diagnosed schizophrenia, the number of patients with a certain amount of mutated genes of the neuronal cytoskeleton is given for different cutoffs (>0 - >7). . .	53
3.17	765 kilo base pair deletion on chromosome 22 displayed in the UCSC Genome Browser [132]. As indicated by the red box, it affects the 22q11 region and is located close to the boundary between 22q11.21 and 22q11.22. In total, 14 genes are affected by the deletion which are annotated below the assembly (blue)	54
4.1	Distribution and frequency of six variant types among schizophrenia samples. For each investigated type, the amount of cases is indicated that is affected by the respective number of events. For <i>de novo</i> knockout (LoF), <i>de novo</i> Variants, Edgetics and ClinVar this event corresponds to the amount of actual SNPs or Indels that were found. For compound heterozygous variants the number indicates the amount of heterozygous variant pairs that were detected as compound heterozygous. In case of the genetic burden analysis, the values indicate the number of genes with increased genetic burden. Subplots for <i>de novo</i> knockout, <i>de novo</i> variants and genetic burden are based on trio cases only. Compound heterozygous, Edgetics and ClinVar include the complete cohort.	69
5.1	(a,b) A neuron with the neuronal cytoskeleton, consisting out of three main structures: microtubules (green), intermediate filaments (purple) and actin filaments (red). (c) The neuronal axon, a membrane-bounded extension, in which microtubule and intermediate filaments form a structural backbone that enables the transport of cargo (e.g. neurotransmitters) from the cell body to the synapse. (d) The growth cone, an actin-supported extension of a developing neurite. (e) A microtubule, formed like a hollow cylinder, consisting out of polymerised α - and β -tubulin dimers. (f) Neurofilaments provide structural support for the axon and regulate the axon diameter. (g) Actin filaments are arranged into networks and play a critical role in neuronal growth and secretion. Figure taken from [184].	76

5.2	Regulation of synaptic glutamate receptors by PDZ proteins. PDZ proteins are characterised by a PSD-95/SAP-90, Discs-large, ZO-1 homologous domain (PDZ) motifs [194]. The synaptic targeting of AMPA receptors requires stargazin binding to PDZ proteins like SAP97/DLG1. In contrast to NMDA receptors, the expression of AMPA receptors is dynamically regulated depending on synapse activity. Figure taken from [194]. . .	79
5.3	Model with Kif13a knockout. With functional 5HT1A receptor transport organised by Kif13a along microtubules, serotonin uptake is maintained at the dendrites. A knockout of Kif13a prevents this transportation of the receptors and consequently the transmission of signals. The released serotonin molecules remain in the synaptic cleft. Figure taken from [198]	80
6.1	Schematic representation of haplotype phasing. By aligning reads against the reference genome, a set of genotypes can be determined. Throughout phasing, the respective haplotypes are extracted. Now, it is possible to call more complex variants like compound heterozygous variants (e.g. Var2 + Var3 and Var2 + Var4)	87

LIST OF ABBREVIATIONS

ALS	Myotrophic Lateral Sclerosis	61
BAM	Binary Alignment/Map	18
CADD	Combined Annotation Dependent Depletion.....	xi
CNV	Copy Number Variation	95
DSM	Diagnostic and Statistical Manual of Mental Disorders	6
GATK	Genome Analysis Toolkit	
GWAS	Genome-wide association studies.....	2
IMPC	International Mouse Phenotyping Consortium.....	32
Indels	Insertions and deletions	16
KNIME	Konstanz Information Miner	32
LD	Linkage Disequilibrium	2
LoF	Loss-of-Function	xii
NCG	Neuronal Cytoskeleton-associated Genes	60
NGS	Next Generation Sequencing.....	16
NIH	National Institute of Mental Health.....	5
PPI	Protein-Protein Interaction	28
rhLoF	Rare Homozygous Loss-of-Function	61
SNP	Single Nucleotide Polymorphism	
VQSR	Variant Quality Score Recalibration	21
22qDS	22q11.2 deletion syndrome	53

BIBLIOGRAPHY

- [1] E. L. Messias, C.-Y. Chen, and W. W. Eaton, “Epidemiology of schizophrenia: review of findings and myths,” *The Psychiatric Clinics of North America*, vol. 30, no. 3, pp. 323–338, Sep. 2007.
- [2] R. Adolphs, “The unsolved problems of neuroscience,” *Trends in Cognitive Sciences*, vol. 19, no. 4, pp. 173–175, Apr. 2015. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1364661315000236>
- [3] H. Fabisch, P. M. Kroisel, and K. Fabisch, “[Genetic risk factors in schizophrenia],” *Fortschritte Der Neurologie-Psychiatrie*, vol. 73 Suppl 1, pp. S44–50, Nov. 2005.
- [4] R. Hilker, D. Helenius, B. Fagerlund, A. Skytthe, K. Christensen, T. M. Werge, M. Nordoft, and B. Glenthøj, “Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register,” *Biological Psychiatry*, Sep. 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0006322317319054>
- [5] R. W. Manderscheid, C. D. Ryff, E. J. Freeman, L. R. McKnight-Eily, S. Dhingra, and T. W. Strine, “Evolving Definitions of Mental Illness and Wellness,” *Preventing Chronic Disease*, vol. 7, no. 1, Dec. 2009. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811514/>
- [6] N. A. on Mental Illness, “Mental Health By the Numbers.” [Online]. Available: <https://www.nami.org/Learn-More/Mental-Health-By-the-Numbers>
- [7] D. Vigo, G. Thornicroft, and R. Atun, “Estimating the true global burden of mental illness,” *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, Feb. 2016. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2215036615005052>
- [8] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove, “The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013,” *International Journal of Epidemiology*, vol. 43, no. 2, pp. 476–493, Apr. 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3997379/>

BIBLIOGRAPHY

- [9] Cross-Disorder Group of the Psychiatric Genomics Consortium, "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *Lancet (London, England)*, vol. 381, no. 9875, pp. 1371–1379, Apr. 2013.
- [10] N. R. Wray, S. H. Lee, and K. S. Kendler, "Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes," *European Journal of Human Genetics*, vol. 20, no. 6, pp. 668–674, Jun. 2012. [Online]. Available: <http://www.nature.com/articles/ejhg2011257>
- [11] E. J. Bromet, R. Kotov, L. J. Fochtmann, G. A. Carlson, M. Tanenberg-Karant, C. Ruggero, and S.-w. Chang, "Diagnostic shifts during the decade following first admission for psychosis," *The American Journal of Psychiatry*, vol. 168, no. 11, pp. 1186–1194, Nov. 2011.
- [12] E. F. Torrey, *Surviving schizophrenia: a manual for families, patients, and providers*, 5th ed. New York: Collins, 2006.
- [13] P. Maki, "Predictors of schizophrenia—a review," *British Medical Bulletin*, vol. 73-74, no. 1, pp. 1–15, Oct. 2005. [Online]. Available: <https://academic.oup.com/bmb/article-lookup/doi/10.1093/bmb/ldh046>
- [14] WHO, "Schizophrenia." [Online]. Available: https://www.who.int/mental_health/management/schizophrenia/en/
- [15] J. van Os and S. Kapur, "Schizophrenia," *The Lancet*, vol. 374, no. 9690, pp. 635–645, Aug. 2009. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0140673609609958>
- [16] J. Kinross, A. Reichenberg, and S. Frangou, "The neurodevelopmental theory of schizophrenia: evidence from studies of early onset cases," *The Israel Journal of Psychiatry and Related Sciences*, vol. 47, no. 2, pp. 110–117, 2010.
- [17] P. C. Sham, C. J. MacLean, and K. S. Kendler, "A typological model of schizophrenia based on age at onset, sex and familial morbidity," *Acta Psychiatrica Scandinavica*, vol. 89, no. 2, pp. 135–141, Feb. 1994.
- [18] A. G. Cardno, E. J. Marshall, B. Coid, A. M. Macdonald, T. R. Ribchester, N. J. Davies, P. Venturi, L. A. Jones, S. W. Lewis, P. C. Sham, I. I. Gottesman, A. E. Farmer, P. McGuffin, A. M. Reveley, and R. M. Murray, "Heritability estimates for psychotic disorders: the Maudsley twin psychosis series," *Archives of General Psychiatry*, vol. 56, no. 2, pp. 162–168, Feb. 1999.

- [19] R. M. Murray, "Mistakes I Have Made in My Research Career," *Schizophrenia Bulletin*, p. sbw165, Dec. 2016. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbw165>
- [20] D. Adam, "Mental health: On the spectrum," *Nature*, vol. 496, no. 7446, pp. 416–418, Apr. 2013. [Online]. Available: <http://www.nature.com/doi/10.1038/496416a>
- [21] M. M. Picchioni and R. M. Murray, "Schizophrenia," *BMJ (Clinical research ed.)*, vol. 335, no. 7610, pp. 91–95, Jul. 2007.
- [22] I. Bombin, "Significance and Meaning of Neurological Signs in Schizophrenia: Two Decades Later," *Schizophrenia Bulletin*, vol. 31, no. 4, pp. 962–977, Sep. 2005. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbi028>
- [23] C. R. Bowie and P. D. Harvey, "Cognitive deficits and functional outcome in schizophrenia," *Neuropsychiatric Disease and Treatment*, vol. 2, no. 4, pp. 531–536, Dec. 2006.
- [24] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition ed. American Psychiatric Association, May 2013. [Online]. Available: <https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596>
- [25] T. Mattila, M. Koeter, T. Wohlfarth, J. Storosum, W. van den Brink, L. de Haan, E. Derks, H. Leufkens, and D. Denys, "Impact of DSM-5 Changes on the Diagnosis and Acute Treatment of Schizophrenia," *Schizophrenia Bulletin*, vol. 41, no. 3, pp. 637–643, May 2015. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbu172>
- [26] A. Ponte, J. Gama Marques, L. Carvalh o Gil, C. Nobrega, S. Pinheiro, and A. Brito, "Catatonic schizophrenia vs anti-NMDA receptor encephalitis – A video case report," *European Psychiatry*, vol. 33, p. S584, Mar. 2016. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0924933816021726>
- [27] L. Lepasavić, I. Lepasavić, B. Šaula-Marojević, and P. Gavrilović, "Paranoid Schizophrenia versus Schizoaffective Disorder: Neuropsychological Aspects," *Srpski Arhiv Za Celokupno Lekarstvo*, vol. 143, no. 7-8, pp. 391–396, Aug. 2015.
- [28] P. Scherzer, E. Leveillé, A. Achim, E. Boisseau, and E. Stip, "A study of theory of mind in paranoid schizophrenia: a theory or many theories?" *Frontiers in Psychology*, vol. 3, p. 432, 2012.
- [29] T. H. McGlashan and W. S. Fenton, "Classical subtypes for schizophrenia: literature review for DSM-IV," *Schizophrenia Bulletin*, vol. 17, no. 4, pp. 609–632, 1991.

BIBLIOGRAPHY

- [30] M. Bengston, *Types of Schizophrenia*, May 2016. [Online]. Available: <https://psychcentral.com/lib/types-of-schizophrenia/>
- [31] W. Hennah, P. Thomson, L. Peltonen, and D. Porteous, "Genes and schizophrenia: beyond schizophrenia: the role of DISC1 in major mental illness," *Schizophrenia Bulletin*, vol. 32, no. 3, pp. 409–416, Jul. 2006.
- [32] S. Gupta and P. Kulhara, "What is schizophrenia: A neurodevelopmental or neurodegenerative disorder or a combination of both? A critical analysis," *Indian Journal of Psychiatry*, vol. 52, no. 1, pp. 21–27, 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824976/>
- [33] T. R. Insel, "Rethinking schizophrenia," *Nature*, vol. 468, no. 7321, pp. 187–193, Nov. 2010. [Online]. Available: <http://www.nature.com/nature/journal/v468/n7321/full/nature09552.html>
- [34] P. Kulhara and S. Gupta, "What is schizophrenia: A neurodevelopmental or neurodegenerative disorder or a combination of both? A critical analysis," *Indian Journal of Psychiatry*, vol. 52, no. 1, p. 21, 2010. [Online]. Available: <http://www.indianjpsychiatry.org/text.asp?2010/52/1/21/58891>
- [35] Y. Zhang and B. R. Bhavnani, "Glutamate-induced apoptosis in neuronal cells is mediated via caspase-dependent and independent mechanisms involving calpain and caspase-3 proteases as well as apoptosis inducing factor (AIF) and this process is inhibited by equine estrogens," *BMC neuroscience*, vol. 7, p. 49, Jun. 2006.
- [36] L. F. Jarskog, L. A. Glantz, J. H. Gilmore, and J. A. Lieberman, "Apoptotic mechanisms in the pathophysiology of schizophrenia," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 29, no. 5, pp. 846–858, Jun. 2005.
- [37] K. L. Davis, R. S. Kahn, G. Ko, and M. Davidson, "Dopamine in schizophrenia: a review and reconceptualization," *The American Journal of Psychiatry*, vol. 148, no. 11, pp. 1474–1486, Nov. 1991.
- [38] O. D. Howes and S. Kapur, "The dopamine hypothesis of schizophrenia: version III—the final common pathway," *Schizophrenia Bulletin*, vol. 35, no. 3, pp. 549–562, May 2009.
- [39] J.-A. Girault and P. Greengard, "The neurobiology of dopamine signaling," *Archives of Neurology*, vol. 61, no. 5, pp. 641–644, May 2004.
- [40] A. Abi-Dargham, O. Mawlawi, I. Lombardo, R. Gil, D. Martinez, Y. Huang, D.-R. Hwang, J. Keilp, L. Kochan, R. Van Heertum, J. M. Gorman, and M. Laruelle, "Prefrontal

- dopamine D1 receptors and working memory in schizophrenia,” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 22, no. 9, pp. 3708–3719, May 2002.
- [41] M. McGue and I. I. Gottesman, “The genetic epidemiology of schizophrenia and the design of linkage studies,” *European Archives of Psychiatry and Clinical Neuroscience*, vol. 240, no. 3, pp. 174–181, Feb. 1991. [Online]. Available: <http://link.springer.com/10.1007/BF02190760>
- [42] K. R. Patel, J. Cherian, K. Gohil, and D. Atkinson, “Schizophrenia: Overview and Treatment Options,” *Pharmacy and Therapeutics*, vol. 39, no. 9, pp. 638–645, Sep. 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159061/>
- [43] T. D. Cannon, J. Kaprio, J. Lönqvist, M. Huttunen, and M. Koskenvuo, “The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study,” *Archives of General Psychiatry*, vol. 55, no. 1, pp. 67–74, Jan. 1998.
- [44] P. F. Sullivan, K. S. Kendler, and M. C. Neale, “Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies,” *Archives of General Psychiatry*, vol. 60, no. 12, pp. 1187–1192, Dec. 2003.
- [45] A. Tenesa and C. S. Haley, “The heritability of human disease: estimation, uses and abuses,” *Nature Reviews Genetics*, vol. 14, no. 2, p. 139, Feb. 2013. [Online]. Available: <https://www.nature.com/articles/nrg3377>
- [46] M. Ng, D. Levinson, S. Faraone, B. Suarez, L. DeLisi, T. Arinami, B. Riley, T. Paunio, A. Pulver, Irmansyah, P. Holmans, M. Escamilla, D. Wildenauer, N. Williams, C. Laurent, B. Mowry, L. Brzustowicz, M. Maziade, P. Sklar, D. Garver, G. Abecasis, B. Lerer, M. Fallin, H. Gurling, P. Gejman, E. Lindholm, H. Moises, W. Byerley, E. Wijsman, P. Forabosco, M. Tsuang, H.-G. Hwu, Y. Okazaki, K. Kendler, B. Wormley, A. Fanous, D. Walsh, F. O’Neill, L. Peltonen, G. Nestadt, V. Lasseter, K. Liang, G. Papadimitriou, D. Dikeos, S. Schwab, M. Owen, M. O’Donovan, N. Norton, E. Hare, H. Raventos, H. Nicolini, M. Albus, W. Maier, V. Nimgaonkar, L. Terenius, J. Mallet, M. Jay, S. Godard, D. Nertney, M. Alexander, R. Crowe, J. Silverman, A. Bassett, M.-A. Roy, C. Mérette, C. Pato, M. Pato, J. L. Roos, Y. Kohn, D. Amann-Zalcenstein, G. Kalsi, A. McQuillin, D. Curtis, J. Brynjolfson, T. Sigmundsson, H. Petursson, A. Sanders, J. Duan, E. Jazin, M. Myles-Worsley, M. Karayiorgou, and C. Lewis, “Meta-analysis of 32 genome-wide linkage studies of schizophrenia,” *Molecular psychiatry*, vol. 14, no. 8, pp. 774–785, Aug. 2009. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2715392/>

BIBLIOGRAPHY

- [47] P. Gejman, A. Sanders, and J. Duan, “The Role of Genetics in the Etiology of Schizophrenia,” *The Psychiatric clinics of North America*, vol. 33, no. 1, pp. 35–66, Mar. 2010. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826121/>
- [48] M. G. Henriksen, J. Nordgaard, and L. B. Jansson, “Genetics of Schizophrenia: Overview of Methods, Findings and Limitations,” *Frontiers in Human Neuroscience*, vol. 11, 2017. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnhum.2017.00322/full>
- [49] B. Xu, J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou, “Exome sequencing supports a de novo mutational paradigm for schizophrenia,” *Nature Genetics*, vol. 43, no. 9, pp. 864–868, Aug. 2011. [Online]. Available: <http://www.nature.com/doifinder/10.1038/ng.902>
- [50] B. Xu, I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. A. Gogos, and M. Karayiorgou, “De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia,” *Nature Genetics*, vol. 44, no. 12, pp. 1365–1369, Dec. 2012. [Online]. Available: <http://www.nature.com/articles/ng.2446>
- [51] S. L. Girard, J. Gauthier, A. Noreau, L. Xiong, S. Zhou, L. Jouan, A. Dionne-Laporte, D. Spiegelman, E. Henrion, O. Diallo, P. Thibodeau, I. Bachand, J. Y. J. Bao, A. H. Y. Tong, C.-H. Lin, B. Millet, N. Jaafari, R. Joobert, P. A. Dion, S. Lok, M.-O. Krebs, and G. A. Rouleau, “Increased exonic de novo mutation rate in individuals with schizophrenia,” *Nature Genetics*, vol. 43, no. 9, pp. 860–863, Sep. 2011. [Online]. Available: <http://www.nature.com/ng/journal/v43/n9/full/ng.886.html>
- [52] P. Jia, X. Chen, A. H. Fanous, and Z. Zhao, “Convergent roles of de novo mutations and common variants in schizophrenia in tissue-specific and spatiotemporal co-expression network,” *Translational Psychiatry*, vol. 8, no. 1, Dec. 2018. [Online]. Available: <http://www.nature.com/articles/s41398-018-0154-2>
- [53] S. E. McCarthy, J. Gillis, M. Kramer, J. Lihm, S. Yoon, Y. Berstein, M. Mistry, P. Pavlidis, R. Solomon, E. Ghiban, E. Antoniou, E. Kelleher, C. O’Brien, G. Donohoe, M. Gill, D. W. Morris, W. R. McCombie, and A. Corvin, “De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability,” *Molecular Psychiatry*, vol. 19, no. 6, pp. 652–658, Jun. 2014. [Online]. Available: <http://www.nature.com/articles/mp201429>
- [54] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari, and

- S. W. Scherer, "Structural Variation of Chromosomes in Autism Spectrum Disorder," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 477–488, Feb. 2008. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0002929707000353>
- [55] S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, Consortium on the Genetics of Schizophrenia (COGS), PAARTNERS Study Group, V. L. Nimgaonkar, R. C. P. Go, R. M. Savage, N. R. Swerdlow, R. E. Gur, D. L. Braff, M.-C. King, and J. M. McClellan, "Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network," *Cell*, vol. 154, no. 3, pp. 518–529, Aug. 2013.
- [56] M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, and M. C. O'Donovan, "De novo mutations in schizophrenia implicate synaptic networks," *Nature*, vol. 506, no. 7487, pp. 179–184, Feb. 2014. [Online]. Available: <http://www.nature.com/articles/nature12929>
- [57] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kähler, L. Duncan, E. Stahl, G. Genovese, E. Fernández, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. E. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. N. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, and P. Sklar, "A polygenic burden of rare disruptive mutations in schizophrenia," *Nature*, vol. 506, no. 7487, pp. 185–190, Feb. 2014. [Online]. Available: <http://www.nature.com/articles/nature12975>
- [58] Swedish Schizophrenia Study, INTERVAL Study, DDD Study, UK10 K Consortium, T. Singh, M. I. Kurki, D. Curtis, S. M. Purcell, L. Crooks, J. McRae, J. Suvisaari, H. Chheda, D. Blackwood, G. Breen, O. Pietiläinen, S. S. Gerety, M. Ayub, M. Blyth, T. Cole, D. Collier, E. L. Coomber, N. Craddock, M. J. Daly, J. Danesh, M. DiForti, A. Foster, N. B. Freimer, D. Geschwind, M. Johnstone, S. Joss, G. Kirov, J. Körkkö, O. Kuismin, P. Holmans, C. M. Hultman, C. Iyegbe, J. Lönnqvist, M. Männikkö, S. A. McCarroll, P. McGuffin, A. M. McIntosh, A. McQuillin, J. S. Moilanen, C. Moore, R. M. Murray, R. Newbury-Ecob, W. Ouwehand, T. Paunio, E. Prigmore, E. Rees, D. Roberts, J. Sambrook, P. Sklar, D. S. Clair, J. Veijola, J. T. R. Walters, H. Williams, P. F. Sullivan, M. E. Hurles, M. C. O'Donovan, A. Palotie, M. J. Owen, and J. C. Barrett, "Rare loss-of-function variants in SETD1a are associated with schizophrenia and

BIBLIOGRAPHY

- developmental disorders,” *Nature Neuroscience*, vol. 19, no. 4, pp. 571–577, Apr. 2016. [Online]. Available: <http://www.nature.com/articles/nn.4267>
- [59] M. B. First, M. Gibbon, R. L. Spitzner, J. B. W. Williams, and L. S. Benjamin, *Structured Clinical Interview for DSM-IV Axis II Personality Disorders*. Arlington VA: American Psychiatric Publishing, Inc., 1997.
- [60] M. B. First, R. L. Spitzner, M. Gibbon, and J. B. W. Williams, *Structured Clinical Interview for DSM-IV Axis I Disorders - Patient Edition (SCID-I/P, Version 2.0)*. New York: New York State Psychiatric Institute, 2002.
- [61] E. J. C. G. Van den Oord, D. Rujescu, J. R. Robles, I. Giegling, C. Birrell, J. Bukszár, L. Murrelle, H.-J. Möller, L. Middleton, and P. Muglia, “Factor structure and external validity of the PANSS revisited,” *Schizophrenia Research*, vol. 82, no. 2-3, pp. 213–223, Feb. 2006.
- [62] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O’Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and Exome Aggregation Consortium, “Analysis of protein-coding genetic variation in 60,706 humans,” *Nature*, vol. 536, no. 7616, pp. 285–291, Aug. 2016. [Online]. Available: <http://www.nature.com/nature/journal/v536/n7616/full/nature19057.html>
- [63] “Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA,” Jul. 2019. [Online]. Available: <https://evs.gs.washington.edu/EVS/>
- [64] The 1000 Genomes Project Consortium, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015. [Online]. Available: <http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html>

- [65] G. Van der Auwera, *Best Practices for Variant Discovery in DNaseq*. [Online]. Available: <https://gatkforums.broadinstitute.org/gatk/discussion/3238/best-practices-for-variant-discovery-in-dnaseq>
- [66] A. E. Renton and B. J. Traynor, "CRESTing the ALS mountain," *Nature Neuroscience*, vol. 16, no. 7, pp. 774–775, Jun. 2013. [Online]. Available: <http://www.nature.com/doi/10.1038/nn.3444>
- [67] M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature Genetics*, vol. 46, no. 3, pp. 310–315, Mar. 2014.
- [68] M. Hastreiter, T. Jeske, J. Hoser, M. Kluge, K. Ahomaa, M.-S. Friedl, S. J. Kopetzky, J.-D. Quell, H. Werner Mewes, and R. Küffner, "KNIME4ngs: a comprehensive toolbox for Next Generation Sequencing analysis," *Bioinformatics*, p. btx003, Jan. 2017. [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx003>
- [69] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.
- [70] *Burrows–Wheeler transform*, Mar. 2017. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Burrows%E2%80%93Wheeler_transform&oldid=768908518
- [71] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008. [Online]. Available: <http://genome.cshlp.org/cgi/doi/10.1101/gr.078212.108>
- [72] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philipakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernysky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics*, vol. 43, no. 5, pp. 491–498, May 2011.
- [73] B. Institute, *Picard Tools - By Broad Institute*, 2009. [Online]. Available: <http://broadinstitute.github.io/picard/>
- [74] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline," *Current Protocols in Bioinformatics*, vol. 43, pp. 11.10.1–33, 2013.

BIBLIOGRAPHY

- [75] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Research*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010. [Online]. Available: <http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110>
- [76] *GATK Doc #4148 HC overview: How the HaplotypeCaller works*. [Online]. Available: <https://software.broadinstitute.org/gatk/documentation/article.php?id=4148>
- [77] A. R. Carson, E. N. Smith, H. Matsui, S. K. Brækkan, K. Jepsen, J.-B. Hansen, and K. A. Frazer, “Effective filtering strategies to improve data quality from population-based whole exome sequencing studies,” *BMC bioinformatics*, vol. 15, p. 125, May 2014.
- [78] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini, “Haplotype Estimation Using Sequencing Reads,” *The American Journal of Human Genetics*, vol. 93, no. 4, pp. 687–696, Oct. 2013. [Online]. Available: [http://www.cell.com/ajhg/abstract/S0002-9297\(13\)00415-1](http://www.cell.com/ajhg/abstract/S0002-9297(13)00415-1)
- [79] G. M. Cooper, E. A. Stone, G. Asimenos, NISC Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow, “Distribution and intensity of constraint in mammalian genomic sequence,” *Genome Research*, vol. 15, no. 7, pp. 901–913, Jul. 2005.
- [80] ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.
- [81] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, “High-resolution mapping and characterization of open chromatin across the genome,” *Cell*, vol. 132, no. 2, pp. 311–322, Jan. 2008.
- [82] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, “Genome-wide mapping of in vivo protein-DNA interactions,” *Science (New York, N.Y.)*, vol. 316, no. 5830, pp. 1497–1502, Jun. 2007.
- [83] P. C. Ng and S. Henikoff, “SIFT: Predicting amino acid changes that affect protein function,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.
- [84] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, “A method and server for predicting damaging missense mutations,” *Nature Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.
- [85] M.-X. Li, H.-S. Gui, J. S. H. Kwan, S.-Y. Bao, and P. C. Sham, “A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases,” *Nucleic Acids Research*, vol. 40, no. 7, p. e53, Apr. 2012.

- [86] M. Funayama, Y. Li, T.-H. Tsoi, C.-W. Lam, T. Ohi, S. Yazawa, E. Uyama, R. Djaldetti, E. Melamed, H. Yoshino, Y. Imamichi, H. Takashima, K. Nishioka, K. Sato, H. Tomiyama, S.-I. Kubo, Y. Mizuno, and N. Hattori, "Familial Parkinsonism with digenic parkin and PINK1 mutations," *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 23, no. 10, pp. 1461–1465, Jul. 2008.
- [87] M. Bolszak, A.-K. Anttonen, T. Komulainen, R. Hinttala, S. Pakanen, R. Sormunen, R. Herva, A.-E. Lehesjoki, K. Majamaa, H. Rantala, and J. Uusimaa, "Digenic mutations in severe myoclonic epilepsy of infancy," *Epilepsy Research*, vol. 85, no. 2-3, pp. 300–304, Aug. 2009.
- [88] S. Girirajan, J. A. Rosenfeld, G. M. Cooper, F. Antonacci, P. Siswara, A. Itsara, L. Vives, T. Walsh, S. E. McCarthy, C. Baker, H. C. Mefford, J. M. Kidd, S. R. Browning, B. L. Browning, D. E. Dickel, D. L. Levy, B. C. Ballif, K. Platky, D. M. Farber, G. C. Gowans, J. J. Wetherbee, A. Asamoah, D. D. Weaver, P. R. Mark, J. Dickerson, B. P. Garg, S. A. Ellingwood, R. Smith, V. C. Banks, W. Smith, M. T. McDonald, J. J. Hoo, B. N. French, C. Hudson, J. P. Johnson, J. R. Ozmores, J. B. Moeschler, U. Surti, L. F. Escobar, D. El-Khechen, J. L. Gorski, J. Kussmann, B. Salbert, Y. Lacassie, A. Biser, D. M. McDonald-McGinn, E. H. Zackai, M. A. Deardorff, T. H. Shaikh, E. Haan, K. L. Friend, M. Fichera, C. Romano, J. Géczy, L. E. DeLisi, J. Sebat, M.-C. King, L. G. Shaffer, and E. E. Eichler, "A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay," *Nature Genetics*, vol. 42, no. 3, pp. 203–209, Mar. 2010.
- [89] V. Harrison, L. Connell, J. Hayesmoore, J. McParland, M. G. Pike, and E. Blair, "Compound heterozygous deletion of NRXN1 causing severe developmental delay with early onset epilepsy in two sisters," *American Journal of Medical Genetics. Part A*, vol. 155A, no. 11, pp. 2826–2831, Nov. 2011.
- [90] D. Roshal, D. Glosser, and A. Zangaladze, "Parieto-occipital lobe epilepsy caused by a POLG1 compound heterozygous A467t/W748s genotype," *Epilepsy & Behavior: E&B*, vol. 21, no. 2, pp. 206–210, Jun. 2011.
- [91] M. Kelly and C. Semsarian, "Multiple Mutations in Genetic Cardiovascular Disease," *Circulation: Cardiovascular Genetics*, vol. 2, no. 2, pp. 182–190, Apr. 2009. [Online]. Available: <http://circgenetics.ahajournals.org/content/2/2/182>
- [92] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The Ensembl Variant Effect Predictor," *Genome Biology*, vol. 17, no. 1, Dec. 2016. [Online]. Available: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4>

BIBLIOGRAPHY

- [93] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009. [Online]. Available: <http://www.nature.com/doi/10.1038/nature08494>
- [94] R. Mosca, J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Céol, M. Soler-López, and P. Aloy, "dSysMap: exploring the edgetic role of disease mutations," *Nature Methods*, vol. 12, no. 3, pp. 167–168, Mar. 2015.
- [95] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, pp. e164–e164, Sep. 2010. [Online]. Available: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603>
- [96] N. Sahni, S. Yi, Q. Zhong, N. Jaikhan, B. Charlotiaux, M. E. Cusick, and M. Vidal, "Edgotype: a fundamental link between genotype and phenotype," *Current Opinion in Genetics & Development*, vol. 23, no. 6, pp. 649–657, Dec. 2013.
- [97] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, no. 3, pp. 281–285, Jul. 1999. [Online]. Available: http://www.nature.com/articles/ng0799_281
- [98] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000. [Online]. Available: http://www.nature.com/articles/ng0500_25
- [99] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, Jan. 1999.
- [100] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Research*, vol. 33, no. Web Server, pp. W741–W748, Jul. 2005. [Online]. Available: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gki475>

- [101] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013,” *Nucleic Acids Research*, vol. 41, no. W1, pp. W77–W83, Jul. 2013. [Online]. Available: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt439>
- [102] Y. H. Benjamini and Y. Hochberg, “Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing,” *ResearchGate*, vol. 57, pp. 289–300, Nov. 1995. [Online]. Available: https://www.researchgate.net/publication/221995234_Controlling_The_False_Discovery_Rate_-_A_Practical_And_Powerful_Approach_To_Multiple_Testing
- [103] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, “STRING v10: protein-protein interaction networks, integrated over the tree of life,” *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D447–452, Jan. 2015.
- [104] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, “Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads,” *Bioinformatics*, vol. 25, no. 21, pp. 2865–2871, Nov. 2009. [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp394>
- [105] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott, “ClinVar: public archive of interpretations of clinically relevant variants,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D862–868, Jan. 2016.
- [106] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen,

S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

- [107] The International Mouse Phenotyping Consortium, M. E. Dickinson, A. M. Flenniken, X. Ji, L. Teboul, M. D. Wong, J. K. White, T. F. Meehan, W. J. Wenginger, H. Westerberg, H. Adissu, C. N. Baker, L. Bower, J. M. Brown, L. B. Caddle, F. Chiani, D. Clary, J. Cleak, M. J. Daly, J. M. Denegre, B. Doe, M. E. Dolan, S. M. Edie, H. Fuchs, V. Gailus-Durner, A. Galli, A. Gambadoro, J. Gallegos, S. Guo, N. R. Horner, C.-W. Hsu, S. J. Johnson, S. Kalaga, L. C. Keith, L. Lanoue, T. N. Lawson, M. Lek, M. Mark, S. Marschall, J. Mason, M. L. McElwee, S. Newbigging, L. M. J. Nutter, K. A. Peterson, R. Ramirez-Solis, D. J. Rowland, E. Ryder, K. E. Samocha, J. R. Seavitt, M. Selloum, Z. Szoke-Kovacs, M. Tamura, A. G. Trainor, I. Tudose, S. Wakana, J. Warren, O. Wendling, D. B. West, L. Wong, A. Yoshiki, W. Wurst, D. G. MacArthur, G. P. Tocchini-Valentini, X. Gao, P. Flicek, A. Bradley, W. C. Skarnes, M. J. Justice, H. E. Parkinson, M. Moore, S. Wells, R. E. Braun, K. L. Svenson, M. H. de Angelis, Y. Herault,

- T. Mohun, A.-M. Mallon, R. M. Henkelman, S. D. M. Brown, D. J. Adams, K. C. K. Lloyd, C. McKerlie, A. L. Beaudet, M. Bućan, and S. A. Murray, “High-throughput discovery of novel developmental phenotypes,” *Nature*, vol. 537, no. 7621, pp. 508–514, Sep. 2016. [Online]. Available: <http://www.nature.com/articles/nature19356>
- [108] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, “KNIME: The Konstanz Information Miner,” in *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 319–326. [Online]. Available: http://link.springer.com/10.1007/978-3-540-78246-9_38
- [109] P. Lindenbaum, S. Le Scouarnec, V. Portero, and R. Redon, “Knime4bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME,” *Bioinformatics*, vol. 27, no. 22, pp. 3200–3201, Nov. 2011. [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr554>
- [110] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, p. R106, 2010. [Online]. Available: <http://dx.doi.org/10.1186/gb-2010-11-10-r106>
- [111] A. Bauch, I. Adamczyk, P. Buczek, F.-J. Elmer, K. Enimanev, P. Glyzowski, M. Kohler, T. Pylak, A. Quandt, C. Ramakrishnan, C. Beisel, L. Malmström, R. Aebersold, and B. Rinn, “openBIS: a flexible framework for managing and analyzing complex data in biology research,” *BMC Bioinformatics*, vol. 12, p. 468, 2011. [Online]. Available: <http://dx.doi.org/10.1186/1471-2105-12-468>
- [112] Schizophrenia Working Group of the Psychiatric Genomics Consortium, “Biological insights from 108 schizophrenia-associated genetic loci,” *Nature*, vol. 511, no. 7510, pp. 421–427, Jul. 2014. [Online]. Available: <http://www.nature.com/articles/nature13595>
- [113] L. Brocchieri, “Protein length in eukaryotic and prokaryotic proteomes,” *Nucleic Acids Research*, vol. 33, no. 10, pp. 3390–3400, Jun. 2005. [Online]. Available: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki615>
- [114] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, and S. Lewis, “AmiGO: online access to ontology and annotation data,” *Bioinformatics*, vol. 25, no. 2, pp. 288–289, Jan. 2009. [Online]. Available: <https://academic.oup.com/bioinformatics/article/25/2/288/220714/AmiGO-online-access-to-ontology-and-annotation>
- [115] A. Büki, G. Kalmár, G. Kekesi, G. Benedek, L. G. Nyúl, and G. Horvath, “Impaired pupillary control in “schizophrenia-like” WISKET rats,” *Autonomic Neuroscience*, vol. 213, pp. 34–42, Sep. 2018. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S156607021730228X>

BIBLIOGRAPHY

- [116] S. E. Robson, M. J. Brookes, E. L. Hall, L. Palaniyappan, J. Kumar, M. Skelton, N. G. Christodoulou, A. Qureshi, F. Jan, M. Z. Katshu, E. B. Liddle, P. F. Liddle, and P. G. Morris, "Abnormal visuomotor processing in schizophrenia," *NeuroImage. Clinical*, vol. 12, pp. 869–878, 2016.
- [117] S. Kivimäe, P.-M. Martin, D. Kapfhamer, Y. Ruan, U. Heberlein, J. L. R. Rubenstein, and B. N. R. Cheyette, "Abnormal behavior in mice mutant for the Disc1 binding partner, Dixdc1," *Translational Psychiatry*, vol. 1, no. 9, pp. e43–e43, Sep. 2011. [Online]. Available: <http://www.nature.com/articles/tp201141>
- [118] A. Putzhammer, B. Heindl, K. Broll, L. Pfeiff, M. Perfahl, and G. Hajak, "Spatial and temporal parameters of gait disturbances in schizophrenic patients," *Schizophrenia Research*, vol. 69, no. 2-3, pp. 159–166, Aug. 2004.
- [119] E. Ozan, E. Yazici, E. Deveci, S. Algul, and I. Kirpinar, "Refusal to eat, as a symptom of schizophrenia, can result in cachexia, phenomenologically resembling comorbid anorexia nervosa," *European Psychiatry*, vol. 23, p. S124, Apr. 2008. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0924933808004446>
- [120] J. Firth, B. Stubbs, D. Vancampfort, J. A. Firth, M. Large, S. Rosenbaum, M. Hallgren, P. B. Ward, J. Sarris, and A. R. Yung, "Grip Strength Is Associated With Cognitive Performance in Schizophrenia and the General Population: A UK Biobank Study of 476559 Participants," *Schizophrenia Bulletin*, vol. 44, no. 4, pp. 728–736, Jun. 2018. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article/44/4/728/4942313>
- [121] V. Kumari, W. Soni, V. M. Mathew, and T. Sharma, "Prepulse inhibition of the startle response in men with schizophrenia: effects of age of onset of illness, symptoms, and medication," *Archives of General Psychiatry*, vol. 57, no. 6, pp. 609–614, Jun. 2000.
- [122] S. J. Akdag, P. G. Nestor, B. F. O'Donnell, M. A. Niznikiewicz, M. E. Shenton, and R. W. McCarley, "The startle reflex in schizophrenia: habituation and personality correlates," *Schizophrenia Research*, vol. 64, no. 2-3, pp. 165–173, Nov. 2003.
- [123] S. Shen, B. Lang, C. Nakamoto, F. Zhang, J. Pu, S.-L. Kuan, C. Chatzi, S. He, I. Mackie, N. J. Brandon, K. L. Marquis, M. Day, O. Hurko, C. D. McCaig, G. Riedel, and D. St Clair, "Schizophrenia-Related Neural and Behavioral Phenotypes in Transgenic Mice Expressing Truncated Disc1," *Journal of Neuroscience*, vol. 28, no. 43, pp. 10 893–10 904, Oct. 2008. [Online]. Available: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3299-08.2008>

- [124] A. Gough and J. Morrison, “Managing the comorbidity of schizophrenia and ADHD,” *Journal of Psychiatry & Neuroscience*, vol. 41, no. 5, pp. E79–E80, Sep. 2016. [Online]. Available: <http://jpn.ca/wp-content/uploads/2016/08/41-5-E79.pdf>
- [125] S. Walther and W. Strik, “Motor symptoms and schizophrenia,” *Neuropsychobiology*, vol. 66, no. 2, pp. 77–92, 2012.
- [126] S. V. Haijma, N. Van Haren, W. Cahn, P. C. M. P. Koolschijn, H. E. Hulshoff Pol, and R. S. Kahn, “Brain Volumes in Schizophrenia: A Meta-Analysis in Over 18 000 Subjects,” *Schizophrenia Bulletin*, vol. 39, no. 5, pp. 1129–1138, Sep. 2013. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbs118>
- [127] A. Vita, L. De Peri, C. Silenzi, and M. Dieci, “Brain morphology in first-episode schizophrenia: A meta-analysis of quantitative magnetic resonance imaging studies,” *Schizophrenia Research*, vol. 82, no. 1, pp. 75–88, Feb. 2006. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0920996405004901>
- [128] R. E. Kaskie, B. Graziano, and F. Ferrarelli, “Schizophrenia and sleep disorders: links, risks, and management challenges,” *Nature and Science of Sleep*, vol. 9, pp. 227–239, 2017.
- [129] T. M. Hyde and D. R. Weinberger, “Seizures and schizophrenia,” *Schizophrenia Bulletin*, vol. 23, no. 4, pp. 611–622, 1997.
- [130] C. Kiran and S. Chaudhury, “Prevalence of comorbid anxiety disorders in schizophrenia,” *Industrial Psychiatry Journal*, vol. 25, no. 1, pp. 35–40, Jun. 2016.
- [131] A. S. Bassett and E. W. Chow, “Schizophrenia and 22q11.2 Deletion Syndrome,” *Current psychiatry reports*, vol. 10, no. 2, pp. 148–157, Apr. 2008. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129332/>
- [132] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler, “The Human Genome Browser at UCSC,” *Genome Research*, vol. 12, no. 6, pp. 996–1006, May 2002. [Online]. Available: <http://www.genome.org/cgi/doi/10.1101/gr.229102>
- [133] P. Gassó, V. Sánchez-Gistau, S. Mas, G. Sugranyes, N. Rodríguez, D. Boloc, E. de la Serna, S. Romero, D. Moreno, C. Moreno, C. M. Díaz-Caneja, A. Lafuente, and J. Castro-Fornieles, “Association of CACNA1c and SYNE1 in offspring of patients with psychiatric disorders,” *Psychiatry Research*, vol. 245, pp. 427–435, Nov. 2016.

BIBLIOGRAPHY

- [134] A. E. Eggers, “A serotonin hypothesis of schizophrenia,” *Medical Hypotheses*, vol. 80, no. 6, pp. 791–794, Jun. 2013. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S030698771300131X>
- [135] H. Y. Meltzer, Z. Li, Y. Kaneda, and J. Ichikawa, “Serotonin receptors: their key role in drugs to treat schizophrenia,” *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 27, no. 7, pp. 1159–1172, Oct. 2003.
- [136] D. M. Ruderfer, E. T. Lim, G. Genovese, J. L. Moran, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, P. Holmans, P. Sklar, and S. M. Purcell, “No evidence for rare recessive and compound heterozygous disruptive variants in schizophrenia,” *European Journal of Human Genetics*, vol. 23, no. 4, pp. 555–557, Apr. 2015. [Online]. Available: <http://www.nature.com/doi/10.1038/ejhg.2014.228>
- [137] M. S. Mostaid, D. Lloyd, B. Liberg, S. Sundram, A. Pereira, C. Pantelis, T. Karl, C. S. Weickert, I. P. Everall, and C. A. Bousman, “Neuregulin-1 and schizophrenia in the genome-wide association study era,” *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 387–409, Sep. 2016. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S014976341630255X>
- [138] Q. Wang, K. He, Z. Li, J. Chen, W. Li, Z. Wen, J. Shen, Y. Qiang, J. Ji, Y. Wang, and Y. Shi, “The CMYA5 gene confers risk for both schizophrenia and major depressive disorder in the Han Chinese population,” *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry*, vol. 15, no. 7, pp. 553–560, Sep. 2014.
- [139] X. Chen, G. Lee, B. S. Maher, A. H. Fanous, J. Chen, Z. Zhao, A. Guo, E. van den Oord, P. F. Sullivan, J. Shi, D. F. Levinson, P. V. Gejman, A. Sanders, J. Duan, M. J. Owen, N. J. Craddock, M. C. O’Donovan, J. Blackman, D. Lewis, G. K. Kirov, W. Qin, S. Schwab, D. Wildenauer, K. Chowdari, V. Nimgaonkar, R. E. Straub, D. R. Weinberger, F. A. O’Neill, D. Walsh, M. Bronstein, A. Darvasi, T. Lencz, A. K. Malhotra, D. Rujescu, I. Giegling, T. Werge, T. Hansen, A. Ingason, M. M. Nöthen, M. Rietschel, S. Cichon, S. Djurovic, O. A. Andreassen, R. M. Cantor, R. Ophoff, A. Corvin, D. W. Morris, M. Gill, C. N. Pato, M. T. Pato, A. Macedo, H. M. D. Gurling, A. McQuillin, J. Pimm, C. Hultman, P. Lichtenstein, P. Sklar, S. M. Purcell, E. Scolnick, D. St Clair, D. H. R. Blackwood, K. S. Kendler, GROUP investigators, and International Schizophrenia Consortium, “GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia,” *Molecular Psychiatry*, vol. 16, no. 11, pp. 1117–1129, Nov. 2011.

- [140] C. A. Ghiani and E. C. Dell'Angelica, "Dysbindin-containing complexes and their proposed functions in brain: from zero to (too) many in a decade," *ASN NEURO*, vol. 3, no. 2, May 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3155195/>
- [141] H. Malkki, "New insights into genetic risk factors for amyotrophic lateral sclerosis: Motor Neuron Disease," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 491–491, Sep. 2016. [Online]. Available: <http://www.nature.com/articles/nrneurol.2016.117>
- [142] R. L. McLaughlin, D. Schijven, W. van Rheenen, K. R. van Eijk, M. O'Brien, R. S. Kahn, R. A. Ophoff, A. Goris, D. G. Bradley, A. Al-Chalabi, L. H. van den Berg, J. J. Luykx, O. Hardiman, J. H. Veldink, A. Shatunov, A. M. Dekker, F. P. Diekstra, S. L. Pulit, R. A. A. van der Spek, P. T. C. van Doormaal, W. Sproviero, A. R. Jones, G. A. Nicholson, D. B. Rowe, R. Pamphlett, M. C. Kiernan, D. Bauer, T. Kahlke, K. Williams, F. Eftimov, I. Fogh, N. Ticozzi, K. Lin, S. Millecamps, F. Salachas, V. Meininger, M. de Carvalho, S. Pinto, J. S. Mora, R. Rojas-García, M. Polak, S. Chandran, S. Colville, R. Swingler, K. E. Morrison, P. J. Shaw, J. Hardy, R. W. Orrell, A. Pittman, K. Sidle, P. Fratta, A. Malaspina, S. Petri, S. Abdulla, C. Drepper, M. Sendtner, T. Meyer, M. Wiedau-Pazos, C. Lomen-Hoerth, V. M. Van Deerlin, J. Q. Trojanowski, L. Elman, L. McCluskey, N. Basak, T. Meitinger, P. Lichtner, M. Blagojevic-Radivojkov, C. R. Andres, C. Maurel, G. Bensimon, B. Landwehrmeyer, A. Brice, C. A. M. Payan, S. Saker-Delye, A. Dürr, N. Wood, L. Tittmann, W. Lieb, A. Franke, M. Rietschel, S. Cichon, M. M. Nöuthen, P. Amouyel, C. Tzourio, J.-F. Dartigues, A. G. Uitterlinden, F. Rivadeneira, K. Estrada, A. Hofman, C. Curtis, A. J. van der Kooi, M. de Visser, M. Weber, C. E. Shaw, B. N. Smith, O. Pansarasa, C. Cereda, R. Del Bo, G. P. Comi, S. D'Alfonso, C. Bertolin, G. Sorarù, L. Mazzini, V. Pensato, C. Gellera, C. Tiloca, A. Ratti, A. Calvo, C. Moglia, M. Brunetti, S. Arcuti, R. Capozzo, C. Zecca, C. Lunetta, S. Penco, N. Riva, A. Padovani, M. Filosto, I. Blair, P. N. Leigh, F. Casale, A. Chio, E. Beghi, E. Pupillo, R. Tortelli, G. Logroscino, J. Powell, A. C. Ludolph, J. H. Weishaupt, W. Robberecht, P. Van Damme, R. H. Brown, J. Glass, J. E. Landers, P. M. Andersen, P. Corcia, P. Vourc'h, V. Silani, M. A. van Es, R. J. Pasterkamp, C. M. Lewis, G. Breen, S. Ripke, B. M. Neale, A. Corvin, J. T. R. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Champion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. K. Chan, R. Y. L. Chan, E. Y. H. Chen, W. Cheng, E. F. C. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell,

J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julià, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kähler, C. Laurent, J. Lee, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lönnqvist, M. Macek, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Meleg, I. Melle, R. I. Meshulam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Müller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietiläinen, J. Pimm, A. J. Pocklington, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H.-C. So, C. C. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Söderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. M. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. O'Donovan, "Genetic correlation between amyotrophic lateral sclerosis and schizophrenia," *Nature Communications*, vol. 8, p. 14774, Mar.

2017. [Online]. Available: <http://www.nature.com/doi/finder/10.1038/ncomms14774>
- [143] I. Bartha, J. d. Iulio, J. C. Venter, and A. Telenti, “Human gene essentiality,” *Nature Reviews Genetics*, vol. 19, no. 1, pp. 51–62, Jan. 2018. [Online]. Available: <https://www.nature.com/articles/nrg.2017.75>
- [144] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, 1000 Genomes Project Consortium, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith, “A systematic survey of loss-of-function variants in human protein-coding genes,” *Science (New York, N.Y.)*, vol. 335, no. 6070, pp. 823–828, Feb. 2012.
- [145] V. M. Narasimhan, K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. R. Barnes, A. H. Barnett, C. Bates, S. Bellary, N. A. Bockett, K. Giorda, C. J. Griffiths, H. Hemingway, Z. Jia, M. A. Kelly, H. A. Khawaja, M. Lek, S. McCarthy, R. McEachan, A. ODonnell-Luria, K. Paigen, C. A. Parisinos, E. Sheridan, L. Southgate, L. Tee, M. Thomas, Y. Xue, M. Schnall-Levin, P. M. Petkov, C. Tyler-Smith, E. R. Maher, R. C. Trembath, D. G. MacArthur, J. Wright, R. Durbin, and D. A. van Heel, “Health and population effects of rare gene knockouts in adult humans with related parents,” *Science*, vol. 352, no. 6284, pp. 474–477, Apr. 2016. [Online]. Available: <http://www.sciencemag.org/cgi/doi/10.1126/science.aac8624>
- [146] Z. Li, Y. Xiang, J. Chen, Q. Li, J. Shen, Y. Liu, W. Li, Q. Xing, Q. Wang, L. Wang, G. Feng, L. He, X. Zhao, and Y. Shi, “Loci with genome-wide associations with schizophrenia in the Han Chinese population,” *British Journal of Psychiatry*, vol. 207, no. 06, pp. 490–494, Dec. 2015. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0007125000239998/type/journal_article
- [147] K. Ohi, T. Shimada, Y. Nitta, H. Kihara, H. Okubo, T. Uehara, and Y. Kawasaki, “Schizophrenia risk variants in ITIH4 and CALN1 regulate gene expression in the dorsolateral prefrontal cortex,” *Psychiatric Genetics*, vol. 26, no. 3, pp. 142–143, Jun. 2016.
- [148] I. S. Consortium, “Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder,” *Nature*, vol. 460, no. 7256, pp. 748–752, Aug. 2009. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912837/>

BIBLIOGRAPHY

- [149] M. R. Salleh, "The genetics of schizophrenia," *The Malaysian journal of medical sciences: MJMS*, vol. 11, no. 2, pp. 3–11, Jul. 2004.
- [150] M. Lin, D. Zhao, A. Hrabovsky, E. Pedrosa, D. Zheng, and H. M. Lachman, "Heat Shock Alters the Expression of Schizophrenia and Autism Candidate Genes in an Induced Pluripotent Stem Cell Model of the Human Telencephalon," *PLoS ONE*, vol. 9, no. 4, Apr. 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3988108/>
- [151] L. Radici, M. Bianchi, R. Crinelli, and M. Magnani, "Ubiquitin C gene: Structure, function, and transcriptional regulation," *Advances in Bioscience and Biotechnology*, vol. 04, no. 12, pp. 1057–1062, 2013. [Online]. Available: <http://www.scirp.org/journal/doi.aspx?DOI=10.4236/abb.2013.412141>
- [152] J. Kaźmierczak-Barańska, Ł. Peczek, P. Przygodzka, and M. J. Cieślak, "Downregulation of striatin leads to hyperphosphorylation of MAP2, induces depolymerization of microtubules and inhibits proliferation of HEK293t cells," *FEBS Letters*, vol. 589, no. 2, pp. 222–230, Jan. 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0014579314008679>
- [153] J. Yan, F. Yan, Z. Li, B. Sinnott, K. M. Cappell, Y. Yu, J. Mo, J. A. Duncan, X. Chen, V. Cormier-Daire, A. W. Whitehurst, and Y. Xiong, "The 3m Complex Maintains Microtubule and Genome Integrity," *Molecular Cell*, vol. 54, no. 5, pp. 791–804, Jun. 2014. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S1097276514002809>
- [154] N. Litterman, Y. Ikeuchi, G. Gallardo, B. C. O'Connell, M. E. Sowa, S. P. Gygi, J. W. Harper, and A. Bonni, "An OBSL1-Cul7fbxw8 Ubiquitin Ligase Signaling Mechanism Regulates Golgi Morphology and Dendrite Patterning," *PLoS Biology*, vol. 9, no. 5, p. e1001060, May 2011. [Online]. Available: <http://dx.plos.org/10.1371/journal.pbio.1001060>
- [155] J. T. M. L. Paridaen, C. Danesin, A. T. Elas, S. van de Water, C. Houart, and D. Zivkovic, "Apc1 is required for maintenance of local brain organizers and dorsal midbrain survival," *Developmental Biology*, vol. 331, no. 2, pp. 101–112, Jul. 2009.
- [156] H. A. Mansour, M. E. Talkowski, J. Wood, K. V. Chowdari, L. McClain, K. Prasad, D. Montrose, A. Fagiolini, E. S. Friedman, M. H. Allen, C. L. Bowden, J. Calabrese, R. S. El-Mallakh, M. Escamilla, S. V. Faraone, M. D. Fossey, L. Gyulai, J. M. Loftis, P. Hauser, T. A. Ketter, L. B. Marangell, D. J. Miklowitz, A. A. Nierenberg, J. Patel, G. S. Sachs, P. Sklar, J. W. Smoller, N. Laird, M. Keshavan, M. E. Thase, D. Axelson, B. Birmaher, D. Lewis, T. Monk, E. Frank, D. J. Kupfer, B. Devlin, and V. L. Nimgaonkar, "Association study of 21 circadian genes with bipolar I disorder, schizoaffective disorder, and schizophrenia," *Bipolar Disorders*, vol. 11, no. 7, pp. 701–710, Nov. 2009.

- [157] T. Bisogno, F. Howell, G. Williams, A. Minassi, M. G. Cascio, A. Ligresti, I. Matias, A. Schiano-Moriello, P. Paul, E.-J. Williams, U. Gangadharan, C. Hobbs, V. Di Marzo, and P. Doherty, "Cloning of the first sn1-DAG lipases points to the spatial and temporal regulation of endocannabinoid signaling in the brain," *The Journal of Cell Biology*, vol. 163, no. 3, pp. 463–468, Nov. 2003.
- [158] S.-H. J. Wang, I. Celic, S.-Y. Choi, M. Riccomagno, Q. Wang, L. O. Sun, S. P. Mitchell, V. Vasioukhin, R. L. Haganir, and A. L. Kolodkin, "Dlg5 regulates dendritic spine formation and synaptogenesis by controlling subcellular N-cadherin localization," *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 34, no. 38, pp. 12 745–12 761, Sep. 2014.
- [159] D. C. F. Lanza, G. V. Meirelles, M. R. Alborghetti, C. H. Abrile, G. Lenz, and J. Kobarg, "FEZ1 interacts with CLASP2 and NEK1 through coiled-coil regions and their cellular colocalization suggests centrosomal functions and regulation by PKC," *Molecular and Cellular Biochemistry*, vol. 338, no. 1-2, pp. 35–45, May 2010.
- [160] M. R. Alborghetti, A. S. Furlan, J. C. Silva, A. F. Paes Leme, I. C. L. Torriani, and J. Kobarg, "Human FEZ1 protein forms a disulfide bond mediated dimer: implications for cargo transport," *Journal of Proteome Research*, vol. 9, no. 9, pp. 4595–4603, Sep. 2010.
- [161] H. M. Knight, B. S. Pickard, A. Maclean, M. P. Malloy, D. C. Soares, A. F. McRae, A. Condie, A. White, W. Hawkins, K. McGhee, M. van Beck, D. J. MacIntyre, J. M. Starr, I. J. Deary, P. M. Visscher, D. J. Porteous, R. E. Cannon, D. St Clair, W. J. Muir, and D. H. R. Blackwood, "A cytogenetic abnormality and rare coding variants identify ABCA13 as a candidate gene in schizophrenia, bipolar disorder, and depression," *American Journal of Human Genetics*, vol. 85, no. 6, pp. 833–846, Dec. 2009.
- [162] N. Jareborg, E. Birney, and R. Durbin, "Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs," *Genome Research*, vol. 9, no. 9, pp. 815–824, Sep. 1999.
- [163] V. Vasiliou, K. Vasiliou, and D. W. Nebert, "Human ATP-binding cassette (ABC) transporter family," *Human Genomics*, vol. 3, no. 3, p. 281, 2008. [Online]. Available: <http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-3-3-281>
- [164] J. M. Gunnarsen, M. H. Kim, S. J. Fuller, M. De Silva, J. M. Britto, V. E. Hammond, P. J. Davies, S. Petrou, E. S. L. Faber, P. Sah, and S.-S. Tan, "Sez-6 proteins affect dendritic arborization patterns and excitability of cortical pyramidal neurons," *Neuron*, vol. 56, no. 4, pp. 621–639, Nov. 2007.

BIBLIOGRAPHY

- [165] L. Van, E. Boot, and A. S. Bassett, "Update on the 22q11.2 deletion syndrome and its relevance to schizophrenia," *Current Opinion in Psychiatry*, vol. 30, no. 3, pp. 191–196, May 2017.
- [166] M. K. Maisenbacher, K. Merrion, B. Pettersen, M. Young, K. Paik, S. Iyengar, S. Kareht, S. Sigurjonsson, Z. P. Demko, and K. A. Martin, "Incidence of the 22q11.2 deletion in a large cohort of miscarriage samples," *Molecular Cytogenetics*, vol. 10, no. 1, Dec. 2017. [Online]. Available: <http://molecularcytogenetics.biomedcentral.com/articles/10.1186/s13039-017-0308-6>
- [167] W. Gao, T. Higaki, M. Eguchi-Ishimae, H. Iwabuki, Z. Wu, E. Yamamoto, H. Takata, M. Ohta, I. Imoto, E. Ishii, and M. Eguchi, "DGCR6 at the proximal part of the DiGeorge critical region is involved in conotruncal heart defects," *Human Genome Variation*, vol. 2, no. 1, Dec. 2015. [Online]. Available: <http://www.nature.com/articles/hgv20154>
- [168] A. Takata, B. Xu, I. Ionita-Laza, J. L. Roos, J. A. Gogos, and M. Karayiorgou, "Loss-of-Function Variants in Schizophrenia Risk and SETD1a as a Candidate Susceptibility Gene," *Neuron*, vol. 82, no. 4, pp. 773–780, May 2014. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0896627314003584>
- [169] J. A. Vorstman, E. W. Chow, R. A. Ophoff, H. van Engeland, F. A. Beemer, R. S. Kahn, R. J. Sinke, and A. S. Bassett, "Association of the PIK4ca schizophrenia-susceptibility gene in adults with the 22q11.2 deletion syndrome," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 150B, no. 3, pp. 430–433, Apr. 2009. [Online]. Available: <http://doi.wiley.com/10.1002/ajmg.b.30827>
- [170] T. G. Nacak, K. Leptien, D. Fellner, H. G. Augustin, and J. Kroll, "The BTB-kelch protein LZTR-1 is a novel Golgi protein that is degraded upon induction of apoptosis," *The Journal of Biological Chemistry*, vol. 281, no. 8, pp. 5065–5071, Feb. 2006.
- [171] J. Mukai, A. Dhillia, L. J. Drew, K. L. Stark, L. Cao, A. B. MacDermott, M. Karayiorgou, and J. A. Gogos, "Palmitoylation-dependent neurodevelopmental deficits in a mouse model of 22q11 microdeletion," *Nature Neuroscience*, vol. 11, no. 11, pp. 1302–1310, Nov. 2008.
- [172] N. Hirokawa, S. Niwa, and Y. Tanaka, "Molecular Motors in Neurons: Transport Mechanisms and Roles in Brain Function, Development, and Disease," *Neuron*, vol. 68, no. 4, pp. 610–638, Nov. 2010. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0896627310007816>
- [173] W.-H. Lin and D. J. Webb, "Actin and Actin-Binding Proteins: Masters of Dendritic Spine Formation, Morphology, and Function," *The Open Neuroscience Journal*, vol. 3, no. 2, pp. 54–66, Dec. 2009. [Online]. Available: <http://www.bentham-open.org/pages/content.php?TONEURJ/2009/00000003/00000002/54TONEURJ.SGM>

- [174] N. J. Bradshaw, V. Bader, I. Prikulis, A. Lueking, S. Müllner, and C. Korth, "Aggregation of the protein TRIOBP-1 and its potential relevance to schizophrenia," *PloS One*, vol. 9, no. 10, p. e111196, 2014.
- [175] J. A. Green and K. Mykytyn, "Neuronal ciliary signaling in homeostasis and disease," *Cellular and Molecular Life Sciences*, vol. 67, no. 19, pp. 3287–3297, Oct. 2010. [Online]. Available: <http://link.springer.com/10.1007/s00018-010-0425-4>
- [176] J. Muñoz-Estrada, A. Lora-Castellanos, I. Meza, S. Alarcón Elizalde, and G. Benítez-King, "Primary cilia formation is diminished in schizophrenia and bipolar disorder: A possible marker for these psychiatric diseases," *Schizophrenia Research*, vol. 195, pp. 412–420, May 2018. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0920996417305303>
- [177] M. C. O'Donovan, N. Craddock, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. A. Spencer, B. Howie, H.-T. Leung, A. M. Hartmann, H.-J. Möller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. M. Williams, I. Giegling, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, P. V. Gejman, Molecular Genetics of Schizophrenia Collaboration, P. V. Gejman, A. R. Sanders, J. Duan, D. F. Levinson, N. G. Buccola, B. J. Mowry, R. Freedman, F. Amin, D. W. Black, J. M. Silverman, W. F. Byerley, C. R. Cloninger, S. Cichon, M. M. Nöthen, M. Gill, A. Corvin, D. Rujescu, G. Kirov, and M. J. Owen, "Identification of loci associated with schizophrenia by genome-wide association and follow-up," *Nature Genetics*, vol. 40, p. 1053, Jul. 2008. [Online]. Available: <https://doi.org/10.1038/ng.201>
- [178] A. B. Kolomeisky, "Motor proteins and molecular motors: how to operate machines at the nanoscale," *Journal of Physics. Condensed Matter: An Institute of Physics Journal*, vol. 25, no. 46, p. 463101, Nov. 2013.
- [179] Q. Xing, R. Gao, H. Li, G. Feng, M. Xu, S. Duan, J. Meng, A. Zhang, S. Qin, and L. He, "Polymorphisms of the *ABCB1* gene are associated with the therapeutic response to risperidone in Chinese schizophrenia patients," *Pharmacogenomics*, vol. 7, no. 7, pp. 987–993, Oct. 2006. [Online]. Available: <https://www.futuremedicine.com/doi/10.2217/14622416.7.7.987>
- [180] M. Pathania, E. C. Davenport, J. Muir, D. F. Sheehan, G. López-Doménech, and J. T. Kittler, "The autism and schizophrenia associated gene *CYFIP1* is critical for the maintenance of dendritic complexity and the stabilization of mature spines," *Translational Psychiatry*, vol. 4, no. 3, pp. e374–e374, Mar. 2014. [Online]. Available: <http://www.nature.com/articles/tp201416>

BIBLIOGRAPHY

- [181] E. Rees, J. T. R. Walters, K. D. Chambert, C. O'Dushlaine, J. Szatkiewicz, A. L. Richards, L. Georgieva, G. Mahoney-Davies, S. E. Legge, J. L. Moran, G. Genovese, D. Levinson, D. W. Morris, P. Cormican, K. S. Kendler, F. A. O'Neill, B. Riley, M. Gill, A. Corvin, Wellcome Trust Case Control Consortium, P. Sklar, C. Hultman, C. Pato, M. Pato, P. F. Sullivan, P. V. Gejman, S. A. McCarroll, M. C. O'Donovan, M. J. Owen, and G. Kirov, "CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1a1 and duplications at 1p36.33 and CGNL1," *Human Molecular Genetics*, vol. 23, no. 6, pp. 1669–1676, Mar. 2014.
- [182] S. Berretta, "Extracellular matrix abnormalities in schizophrenia," *Neuropharmacology*, vol. 62, no. 3, pp. 1584–1597, Mar. 2012. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0028390811003431>
- [183] J. Su, J. Chen, K. Lippold, A. Monavarfeshani, G. L. Carrillo, R. Jenkins, and M. A. Fox, "Collagen-derived matricryptins promote inhibitory nerve terminal formation in the developing neocortex," *The Journal of Cell Biology*, vol. 212, no. 6, pp. 721–736, Mar. 2016. [Online]. Available: <http://www.jcb.org/lookup/doi/10.1083/jcb.201509085>
- [184] D. A. Fletcher and R. D. Mullins, "Cell mechanics and the cytoskeleton," *Nature*, vol. 463, no. 7280, pp. 485–492, Jan. 2010. [Online]. Available: <http://www.nature.com/doi/10.1038/nature08908>
- [185] M. Kuijpers and C. C. Hoogenraad, "Centrosomes, microtubules and neuronal development," *Molecular and Cellular Neuroscience*, vol. 48, no. 4, pp. 349–358, Dec. 2011. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S1044743111001102>
- [186] K. D. Meyer and J. A. Morris, "Disc1 regulates granule cell migration in the developing hippocampus," *Human Molecular Genetics*, vol. 18, no. 17, pp. 3286–3297, Sep. 2009. [Online]. Available: <http://www.hmg.oxfordjournals.org/cgi/doi/10.1093/hmg/ddp266>
- [187] L. C. Kapitein and C. C. Hoogenraad, "Building the Neuronal Microtubule Cytoskeleton," *Neuron*, vol. 87, no. 3, pp. 492–506, Aug. 2015. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0896627315005139>
- [188] P. A. Coulombe, O. Bousquet, L. Ma, S. Yamada, and D. Wirtz, "The 'ins' and 'outs' of intermediate filament organization," *Trends in Cell Biology*, vol. 10, no. 10, pp. 420–428, Oct. 2000.
- [189] J. Kushkuley, W. K. H. Chan, S. Lee, J. Eyer, J.-F. Leterrier, F. Letournel, and T. B. Shea, "Neurofilament cross-bridging competes with kinesin-dependent association of neurofilaments with microtubules," *Journal of Cell Science*, vol. 122, no. 19, pp. 3579–3586, Oct. 2009. [Online]. Available: <http://jcs.biologists.org/cgi/doi/10.1242/jcs.051318>

- [190] H. Wang, M. Wu, C. Zhan, E. Ma, M. Yang, X. Yang, and Y. Li, “Neurofilament proteins in axonal regeneration and neurodegenerative diseases,” *Neural Regeneration Research*, vol. 7, no. 8, pp. 620–626, Mar. 2012. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4346988/>
- [191] Y. Liu, C. Gervasi, and B. G. Szaro, “A crucial role for hnRNP K in axon development in *Xenopus laevis*,” *Development*, vol. 135, no. 18, pp. 3125–3135, Sep. 2008. [Online]. Available: <http://dev.biologists.org/cgi/doi/10.1242/dev.022236>
- [192] T. Shirao and C. González-Billault, “Actin filaments and microtubules in dendritic spines,” *Journal of Neurochemistry*, vol. 126, no. 2, pp. 155–164, Jul. 2013. [Online]. Available: <http://doi.wiley.com/10.1111/jnc.12313>
- [193] M. A. Gomez-Ferreria, M. Bashkurov, M. Mullin, A.-C. Gingras, and L. Pelletier, “CEP192 interacts physically and functionally with the K63-deubiquitinase CYLD to promote mitotic spindle assembly,” *Cell Cycle*, vol. 11, no. 19, pp. 3555–3558, Oct. 2012. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.4161/cc.21574>
- [194] S. Tomita, R. A. Nicoll, and D. S. Brecht, “PDZ Protein Interactions Regulating Glutamate Receptor Function and Plasticity,” *The Journal of Cell Biology*, vol. 153, no. 5, pp. F19–F24, May 2001. [Online]. Available: <http://www.jcb.org/lookup/doi/10.1083/jcb.153.5.F19>
- [195] A. Uezato, N. Yamamoto, Y. Iwayama, S. Hiraoka, E. Hiraaki, A. Umino, E. Haramo, M. Umino, T. Yoshikawa, and T. Nishikawa, “Reduced cortical expression of a newly identified splicing variant of the DLG1 gene in patients with early-onset schizophrenia,” *Translational Psychiatry*, vol. 5, no. 10, p. e654, Oct. 2015. [Online]. Available: <http://www.nature.com/doi/10.1038/tp.2015.154>
- [196] J.-B. Manneville, M. Jehanno, and S. Etienne-Manneville, “Dlg1 binds GKAP to control dynein association with microtubules, centrosome positioning, and cell polarity,” *The Journal of Cell Biology*, vol. 191, no. 3, pp. 585–598, Nov. 2010. [Online]. Available: <http://www.jcb.org/lookup/doi/10.1083/jcb.201002151>
- [197] Y. Vyas and J. Montgomery, “The role of postsynaptic density proteins in neural degeneration and regeneration,” *Neural Regeneration Research*, vol. 11, no. 6, p. 0, 2016. [Online]. Available: <http://www.nrronline.org/text.asp?2016/11/6/0/184481>
- [198] R. Zhou, S. Niwa, L. Guillaud, Y. Tong, and N. Hirokawa, “A Molecular Motor, KIF13a, Controls Anxiety by Transporting the Serotonin Type 1a Receptor,” *Cell Reports*, vol. 3, no. 2, pp. 509–519, Feb. 2013. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S2211124713000211>

BIBLIOGRAPHY

- [199] K. H. Yamada, Y. Nakajima, M. Geyer, K. K. Wary, M. Ushio-Fukai, Y. Komarova, and A. B. Malik, “KIF13b regulates angiogenesis through Golgi to plasma membrane trafficking of VEGFR2,” *Journal of Cell Science*, vol. 127, no. 20, pp. 4518–4530, Oct. 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4197089/>
- [200] A. Bellon, J. Luchino, K. Haigh, G. Rougon, J. Haigh, S. Chauvet, and F. Mann, “VEGFR2 (KDR/Flk1) signaling mediates axon growth in response to semaphorin 3e in the developing brain,” *Neuron*, vol. 66, no. 2, pp. 205–219, Apr. 2010.
- [201] S. M. Guadiana and E. A. Grove, “Cilia on the Brain: Primary Cilia and Their Roles in Brain Function,” in *eLS*, John Wiley & Sons Ltd, Ed. Chichester, UK: John Wiley & Sons, Ltd, Feb. 2015, pp. 1–11. [Online]. Available: <http://doi.wiley.com/10.1002/9780470015902.a0025793>
- [202] A. Marley and M. von Zastrow, “DISC1 Regulates Primary Cilia That Display Specific Dopamine Receptors,” *PLoS ONE*, vol. 5, no. 5, p. e10902, May 2010. [Online]. Available: <http://dx.plos.org/10.1371/journal.pone.0010902>
- [203] D. van de Willige, C. C. Hoogenraad, and A. Akhmanova, “Microtubule plus-end tracking proteins in neuronal development,” *Cellular and Molecular Life Sciences*, vol. 73, pp. 2053–2077, 2016. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4834103/>
- [204] S. D. Ryan, K. Bhanot, A. Ferrier, Y. De Repentigny, A. Chu, A. Blais, and R. Kothary, “Microtubule stability, Golgi organization, and transport flux require dystonin-a2-MAP1b interaction,” *The Journal of Cell Biology*, vol. 196, no. 6, pp. 727–742, Mar. 2012. [Online]. Available: <http://www.jcb.org/lookup/doi/10.1083/jcb.201107096>
- [205] E. M. Kenny, P. Cormican, S. Furlong, E. Heron, G. Kenny, C. Fahey, E. Kelleher, S. Ennis, D. Tropea, R. Anney, A. P. Corvin, G. Donohoe, L. Gallagher, M. Gill, and D. W. Morris, “Excess of rare novel loss-of-function variants in synaptic genes in schizophrenia and autism spectrum disorders,” *Molecular Psychiatry*, vol. 19, no. 8, pp. 872–879, Aug. 2014. [Online]. Available: <http://www.nature.com/articles/mp2013127>
- [206] J. Costas, J. J. Suárez-Rama, N. Carrera, E. Paz, M. Páramo, S. Agra, J. Brenlla, R. Ramos-Ríos, and M. Arrojo, “Role of DISC1 interacting proteins in schizophrenia risk from genome-wide analysis of missense SNPs,” *Annals of Human Genetics*, vol. 77, no. 6, pp. 504–512, Nov. 2013.
- [207] J. Guo, H. Higginbotham, J. Li, J. Nichols, J. Hirt, V. Ghukasyan, and E. Anton, “Developmental disruptions underlying brain abnormalities in ciliopathies,” *Nature Communications*, vol. 6, Jul. 2015. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4515781/>

- [208] H. H. Arts, D. Doherty, S. E. C. van Beersum, M. A. Parisi, S. J. F. Letteboer, N. T. Gorden, T. A. Peters, T. Märker, K. Voeselek, A. Kartono, H. Ozyurek, F. M. Farin, H. Y. Kroes, U. Wolfrum, H. G. Brunner, F. P. M. Cremers, I. A. Glass, N. V. A. M. Knoers, and R. Roepman, “Mutations in the gene encoding the basal body protein RPGRIP11, a nephrocystin-4 interactor, cause Joubert syndrome,” *Nature Genetics*, vol. 39, no. 7, pp. 882–888, Jul. 2007. [Online]. Available: <http://www.nature.com/articles/ng2069>
- [209] S. Thakurela, N. Tiwari, S. Schick, A. Garding, R. Ivanek, B. Berninger, and V. K. Tiwari, “Mapping gene regulatory circuitry of Pax6 during neurogenesis,” *ResearchGate*, vol. 2, p. 15045, Feb. 2016. [Online]. Available: https://www.researchgate.net/publication/293808747_Mapping_gene_regulatory_circuitry_of_Pax6_during_neurogenesis
- [210] S. M. Sisodiya, S. L. Free, K. A. Williamson, T. N. Mitchell, C. Willis, J. M. Stevens, B. E. Kendall, S. D. Shorvon, I. M. Hanson, A. T. Moore, and V. van Heyningen, “PAX6 haploinsufficiency causes cerebral malformation and olfactory dysfunction in humans,” *Nature Genetics*, vol. 28, no. 3, pp. 214–216, Jul. 2001.
- [211] G. Stöber, Y. V. Syagailo, O. Okladnova, G. Jungkunz, M. Knapp, H. Beckmann, and K. P. Lesch, “Functional PAX-6 gene-linked polymorphic region: potential association with paranoid schizophrenia,” *Biological Psychiatry*, vol. 45, no. 12, pp. 1585–1591, Jun. 1999.
- [212] T. Sapir, T. Levy, A. Sakakibara, A. Rabinkov, T. Miyata, and O. Reiner, “Shootin1 acts in concert with KIF20b to promote polarization of migrating neurons,” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 33, no. 29, pp. 11 932–11 948, Jul. 2013.
- [213] C. Gao, S. F. Frausto, A. L. Guedea, N. C. Tronson, V. Jovasevic, K. Leaderbrand, K. A. Corcoran, Y. F. Guzmán, G. T. Swanson, and J. Radulovic, “IQGAP1 regulates NR2a signaling, spine density, and cognitive processes,” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 31, no. 23, pp. 8533–8542, Jun. 2011. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3121195/>
- [214] M. A. Snyder and W.-J. Gao, “NMDA hypofunction as a convergence point for progression and symptoms of schizophrenia,” *Frontiers in Cellular Neuroscience*, vol. 7, Mar. 2013. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3608949/>
- [215] B. J. O’Roak, P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. Mackenzie, S. B. Ng, C. Baker, M. J. Rieder, D. A. Nickerson, R. Bernier, S. E. Fisher, J. Shendure, and E. E. Eichler, “Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations,” *Nature Genetics*, vol. 43, no. 6, pp. 585–589, Jun. 2011.

BIBLIOGRAPHY

- [216] B. I. Drögemöller, R. Emsley, B. Chiliza, L. van der Merwe, G. E. B. Wright, M. Daya, E. Hoal, A. K. Malhotra, T. Lencz, D. G. Robinson, J.-P. Zhang, L. Asmal, D. J. H. Niehaus, and L. Warnich, “The identification of novel genetic variants associated with antipsychotic treatment response outcomes in first-episode schizophrenia patients,” *Pharmacogenetics and Genomics*, vol. 26, no. 5, pp. 235–242, May 2016.
- [217] N. Gogtay, N. S. Vyas, R. Testa, S. J. Wood, and C. Pantelis, “Age of Onset of Schizophrenia: Perspectives From Structural Neuroimaging Studies,” *Schizophrenia Bulletin*, vol. 37, no. 3, pp. 504–513, May 2011. [Online]. Available: <https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/sbr030>
- [218] M. T. W. Ebbert, M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, J. Duce, Alzheimer’s Disease Neuroimaging Initiative, J. S. K. Kauwe, and P. G. Ridge, “Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches,” *BMC bioinformatics*, vol. 17 Suppl 7, p. 239, Jul. 2016.
- [219] A. Sekar, A. R. Bialas, H. de Rivera, A. Davis, T. R. Hammond, N. Kamitaki, K. Tooley, J. Presumey, M. Baum, V. Van Doren, G. Genovese, S. A. Rose, R. E. Handsaker, M. J. Daly, M. C. Carroll, B. Stevens, and S. A. McCarroll, “Schizophrenia risk from complex variation of complement component 4,” *Nature*, vol. 530, no. 7589, pp. 177–183, Jan. 2016. [Online]. Available: <http://www.nature.com/doi/10.1038/nature16549>
- [220] N. Altman and M. Krzywinski, “Association, correlation and causation,” *Nature Methods*, vol. 12, p. 899, Sep. 2015. [Online]. Available: <http://dx.doi.org/10.1038/nmeth.3587>
- [221] E. C. Johnson, R. Border, W. E. Melroy-Greif, C. A. de Leeuw, M. A. Ehringer, and M. C. Keller, “No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes,” *Biological Psychiatry*, vol. 82, no. 10, pp. 702–708, Nov. 2017. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0006322317317729>
- [222] M. S. Farrell, T. Werge, P. Sklar, M. J. Owen, R. A. Ophoff, M. C. O’Donovan, A. Corvin, S. Cichon, and P. F. Sullivan, “Evaluating historical candidate genes for schizophrenia,” *Molecular Psychiatry*, vol. 20, no. 5, pp. 555–562, May 2015. [Online]. Available: <http://www.nature.com/articles/mp201516>
- [223] P. Jia, G. Han, J. Zhao, P. Lu, and Z. Zhao, “SZGR 2.0: a one-stop shop of schizophrenia candidate genes,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D915–D924, 2017.
- [224] G. Benitez-King, G. Ramírez-Rodríguez, L. Ortíz, and I. Meza, “The neuronal cytoskeleton as a potential therapeutic target in neurodegenerative diseases and schizophrenia,” *Current drug targets. CNS and neurological disorders*, vol. 3, no. 6, pp. 515–533, Dec. 2004.

- [225] L. R. Gabriel, S. Wu, P. Kearney, K. D. Bellvé, C. Standley, K. E. Fogarty, and H. E. Melikian, “Dopamine transporter endocytic trafficking in striatal dopaminergic neurons: differential dependence on dynamin and the actin cytoskeleton,” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 33, no. 45, pp. 17 836–17 846, Nov. 2013.
- [226] A. Y. Park, Y. S. Park, D. So, I.-K. Song, J.-E. Choi, H.-J. Kim, and K.-J. Lee, “Activity-Regulated Cytoskeleton-Associated Protein (Arc/Arg3.1) is Transiently Expressed after Heat Shock Stress and Suppresses Heat Shock Factor 1,” *Scientific Reports*, vol. 9, no. 1, Dec. 2019. [Online]. Available: <http://www.nature.com/articles/s41598-019-39292-1>
- [227] L. C. Kapitein, K. W. Yau, S. M. Gouveia, W. A. van der Zwan, P. S. Wulf, N. Keijzer, J. Demmers, J. Jaworski, A. Akhmanova, and C. C. Hoogenraad, “NMDA Receptor Activation Suppresses Microtubule Growth and Spine Entry,” *Journal of Neuroscience*, vol. 31, no. 22, pp. 8194–8209, Jun. 2011. [Online]. Available: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.6215-10.2011>
- [228] C. R. Marshall, D. P. Howrigan, D. Merico, B. Thiruvahindrapuram, W. Wu, D. S. Greer, D. Antaki, A. Shetty, P. A. Holmans, D. Pinto, M. Gujral, W. M. Brandler, D. Malhotra, Z. Wang, K. V. F. Fajardo, M. S. Maile, S. Ripke, I. Agartz, M. Albus, M. Alexander, F. Amin, J. Atkins, S. A. Bacanu, R. A. Belliveau Jr, S. E. Bergen, M. Bertalan, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, B. Bulik-Sullivan, W. Byerley, W. Cahn, G. Cai, M. J. Cairns, D. Champion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, W. Cheng, C. R. Cloninger, D. Cohen, P. Cormican, N. Craddock, B. Crespo-Facorro, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, K.-H. Farh, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, J. I. Friedman, A. J. Forstner, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, J. Gratten, L. de Haan, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, H. Huang, M. Ikeda, I. Joa, A. K. Kähler, R. S. Kahn, L. Kalaydjieva, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, Y. Kim, J. A. Knowles, B. Konte, C. Laurent, P. Lee, S. H. Lee, S. E. Legge, B. Lerer, D. L. Levy, K.-Y. Liang, J. Lieberman, J. Lönngqvist, C. M. Loughland, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, I. Melle, R. I. Meshulam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrabb, D. W. Morris, B. Müller-Myhsok, K. C. Murphy, R. M. Murray, I. Myin-Germeys, I. Nenadic,

- D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nisenbaum, A. Nordin, E. O’Callaghan, C. O’Dushlaine, S.-Y. Oh, A. Olincy, L. Olsen, F. A. O’Neill, J. Van Os, C. Pantelis, G. N. Papadimitriou, E. Parkhomenko, M. T. Pato, T. Paunio, Psychosis Endophenotypes International Consortium, D. O. Perkins, T. H. Pers, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, A. Savitz, U. Schall, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, J. M. Silverman, J. W. Smoller, E. Söderman, C. C. A. Spencer, E. A. Stahl, E. Strengman, J. Strohmaier, T. S. Stroup, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, S. Thirumalai, P. A. Tooney, J. Veijola, P. M. Visscher, J. Waddington, D. Walsh, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, B. K. Wormley, N. R. Wray, J. Q. Wu, C. C. Zai, R. Adolfsson, O. A. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, S. Cichon, D. A. Collier, A. Corvin, M. J. Daly, A. Darvasi, E. Domenici, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, D. F. Levinson, Q. S. Li, S. A. McCarroll, A. McQuillin, J. L. Moran, B. J. Mowry, M. M. Nöthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. Sklar, D. St Clair, J. T. R. Walters, T. Werge, P. F. Sullivan, M. C. O’Donovan, S. W. Scherer, B. M. Neale, J. Sebat, and CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium, “Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects,” *Nature Genetics*, vol. 49, p. 27, Nov. 2016. [Online]. Available: <http://dx.doi.org/10.1038/ng.3725>
- [229] B. S. Gloss and M. E. Dinger, “Realizing the significance of noncoding functionality in clinical genomics,” *Experimental & Molecular Medicine*, vol. 50, no. 8, Aug. 2018. [Online]. Available: <http://www.nature.com/articles/s12276-018-0087-0>
- [230] J. Zhou, C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel, J. J. Fak, J. Funk, K. Yao, Y. Tajima, A. Packer, R. B. Darnell, and O. G. Troyanskaya, “Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk,” *Nature Genetics*, vol. 51, no. 6, pp. 973–980, Jun. 2019. [Online]. Available: <http://www.nature.com/articles/s41588-019-0420-0>
- [231] F. Pfeiffer, C. Gröber, M. Blank, K. Händler, M. Beyer, J. L. Schultze, and G. Mayer, “Systematic evaluation of error rates and causes in short samples in next-generation sequencing,” *Scientific Reports*, vol. 8, no. 1, Dec. 2018. [Online]. Available: <http://www.nature.com/articles/s41598-018-29325-6>
- [232] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani, “Estimation of sequencing error rates in short reads,” *BMC bioinformatics*, vol. 13, p. 185, Jul. 2012.

- [233] J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson, “Induced pluripotent stem cell lines derived from human somatic cells,” *Science (New York, N.Y.)*, vol. 318, no. 5858, pp. 1917–1920, Dec. 2007.
- [234] Z. P. Pang, N. Yang, T. Vierbuchen, A. Ostermeier, D. R. Fuentes, T. Q. Yang, A. Citri, V. Sebastiano, S. Marro, T. C. Südhof, and M. Wernig, “Induction of human neuronal cells by defined transcription factors,” *Nature*, vol. 476, no. 7359, pp. 220–223, May 2011.
- [235] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, “Multiplex genome engineering using CRISPR/Cas systems,” *Science (New York, N.Y.)*, vol. 339, no. 6121, pp. 819–823, Feb. 2013.
- [236] K. Takahashi and S. Yamanaka, “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors,” *Cell*, vol. 126, no. 4, pp. 663–676, Aug. 2006.
- [237] T. Vierbuchen, A. Ostermeier, Z. P. Pang, Y. Kokubu, T. C. Südhof, and M. Wernig, “Direct conversion of fibroblasts to functional neurons by defined factors,” *Nature*, vol. 463, no. 7284, pp. 1035–1041, Feb. 2010.
- [238] L. Muzio and G. G. Consalez, “Modeling human brain development with cerebral organoids,” *Stem Cell Research & Therapy*, vol. 4, no. 6, p. 154, 2013.

