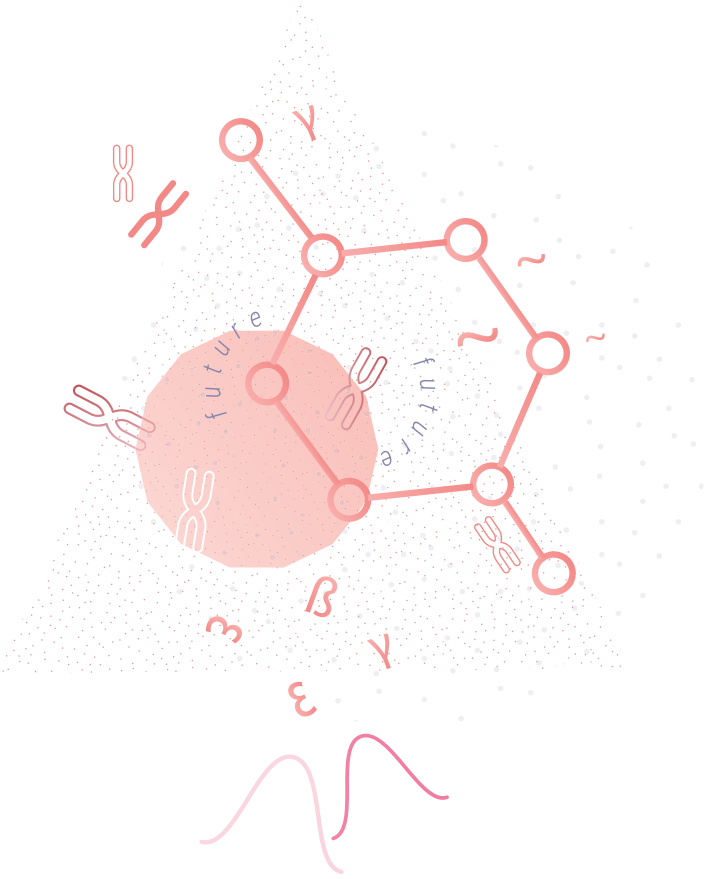
From genetic variants to biological pathways in neuropsychiatric traits

Hammerschlag, A.R.

2018

document version
Publisher's PDF, also known as Version of record

Chapter 7



Summary, discussion and future directions

This thesis contains a collection of genetic studies that address different neuropsychiatric traits and use different analytic approaches. The main objective of all studies was to identify genetic variants, genes or biological pathways for neuropsychiatric traits. The different chapters of my thesis illustrate the rapid advances made in the research field of neuropsychiatric genetics. In this final chapter, I summarize the findings of my research, followed by a general discussion of several topics that were central in my thesis and future directions in this research field.

1. Summary

The first part of this thesis describes two different methods to identify genetic variants, genes and biological pathways for Attention-Deficit/Hyperactivity Disorder (ADHD). In **chapter 2**, we investigated ADHD symptom scores in a large meta-analysis of genome-wide association studies (GWAS) in population-based pediatric cohorts. Analyzing these cohorts allowed us to study ADHD related behaviors in a much larger sample than the available samples containing ADHD diagnosed individuals. It has been suggested that a diagnosis of ADHD can be regarded as the extreme end of a continuous distribution of inattentive and hyperactive behaviors. Indeed, we showed that ADHD symptom scores and an ADHD diagnosis have strong genetic overlap (r_g = 0.96), suggesting that the same genetic factors underlie both phenotypes. Our GWAS of 17,666 children did not result in genome-wide significant loci, but we did detect three gene-wide significant genes, of which one is involved in neuronal development. Combining data from population-based and clinical cohorts to increase sample size and improve statistical power for GWAS will be a next step to identifying genetic variants for ADHD.

In **chapter 3**, we applied a different approach to test the hypothesis that synaptic function is involved in ADHD etiology. We performed a functional gene-set analysis to test whether synaptic processes are associated with ADHD. Gene-set analysis tests the joint effect

of multiple genetic variants in groups of functionally related genes, in order to increase statistical power compared to conventional GWAS. Nevertheless, we did not identify specific synaptic function categories for ADHD. Given the sample size of the study and the gene sets based on the biological knowledge at that time, our results do not support a large effect of common genetic variants in synaptic genes on ADHD. Larger sample sizes will be required to study a possible role of synaptic genes with smaller effects on ADHD.

We continued with investigating additional psychiatric disorders, with a focus on the genetic overlap between different disorders, which I describe in chapter 4. Because mounting evidence shows genetic overlap between multiple psychiatric disorders, we aimed to identify biological pathways that can explain this shared risk. By applying gene-set analysis combining data on five psychiatric disorders (schizophrenia, bipolar disorder, major depressive disorder, autism spectrum disorder (ASD), and ADHD), we found postsynaptic, neuron projection and brain-expressed sets of genes. Interestingly, not only the adult-onset disorders showed overlap in their associations with these sets, the childhood-onset disorders contributed to part of the gene-set associations as well. This suggests that specific biological functions play a role in the etiology of several psychiatric disorders across the life span.

The previous three studies utilized data of common genetic variants to investigate the etiology of neuropsychiatric disorders. However, rare genetic variants have been suggested to play an additional role. In chapter 5 we investigated the contribution of rare coding variants to the genetic architecture of alcohol consumption and tobacco use using data from the ExomeChip from eight research groups. We were not able to identify genome-wide significant rare variants in single-variant and gene-based tests for alcohol consumption and number of cigarettes per day despite large sample sizes of 12,466 and 7,432 individuals, respectively. Our results imply that rare variants in the exome with large effects do not contribute to variance in the two substance use traits.

In the final study of this thesis described in chapter 6 we performed a GWAS on a large sample of 113,006 individuals. The GWAS and gene-based analysis revealed several novel loci and genes for insomnia complaints, with *MEIS1* showing the strongest association. *MEIS1* has been related to restless legs syndrome (RLS) before and we showed a likely pleiotropic effect of *MEIS1* on both conditions. Sex-specific analyses suggested sex-specific as well as shared genetic factors across males and females. Furthermore, we showed substantial positive genetic overlap with internalizing and metabolic traits and a negative overlap with subjective well-being and educational attainment. These findings provide novel insight into the genetic architecture of insomnia complaints.

2. General Discussion and future perspectives

2.1 Increase statistical power for GWAS – Sample size does matter

In **chapters 2 and 3** we investigated the genetics of ADHD. Apart from the three genes that we identified in **chapter 2**, we could not identify additional genetic variants, genes or biological pathways for ADHD. Most likely this results from insufficient statistical power to detect the effects of the genetic variants involved, even when we aggregate their effects in gene sets, as we report in **chapter 3**. ADHD is a polygenic trait, i.e. each individual will carry multiple alleles that increase the risk to develop ADHD, and multiple alleles that decrease this risk. Each individual variant will typically explain only a very small proportion of the variance in the phenotype. In addition, because of the many potentially different combinations of these risk alleles, it is likely that each individual carries a unique set. In GWAS, the effect size of each allele is measured against an averaged background, which makes the measured effect sizes of these alleles small and difficult to detect. The observations in my thesis imply that much larger sample sizes are needed to be able to pick up the small signals.

Despite the few findings in the GWAS meta-analysis for ADHD symptom scores in **chapter 2**, the study was an important step in the search for the genetic underpinnings of ADHD. The ADHD symptom scores had a strikingly high genetic correlation with ADHD based on diagnostic criteria (r_g=0.96) and polygenic risk scores based on symptom scores highly predict ADHD status in clinical samples[1–3]. These findings suggest that the instruments to score ADHD symptoms assess an underlying common liability for ADHD, and support the hypothesis that ADHD might be the extreme of these quantitative traits. Hence, we proposed to combine continuous ADHD measures assessed in the general population with dichotomous diagnosis of ADHD assessed in clinical samples to increase sample size and statistical power in future GWAS. This design was applied in a follow up study by Demontis and colleagues[4] where the ADHD symptom scores GWAS was meta-analyzed with the clinical sample of the Psychiatric Genomics Consortium (PGC) which resulted in a sample of 149,290 individuals. Indeed, with this sample size the first 16 independent genomic loci associated with ADHD were identified. The majority of these significantly associated loci showed a concordant direction of effect between the two samples. The results indicate that genes involved in neurodevelopmental processes, including *FOXP2* and *DUSP6*, play an important role in ADHD risk. Furthermore, the hypothesized polygenic architecture of ADHD was supported by these findings; only a tiny fraction of common variant risk for ADHD could be explained by the identified loci, and the effect sizes of the loci were modest with odds ratios (OR) ranging from 1.08 to 1.20. When considering *all* common variants included in the GWAS, a considerable amount of the heritability of ADHD could be explained (h^2_{SNP} = 0.22, SE = 0.01), which is a typical observation for polygenic psychiatric disorders.

We can conclude not only from these ADHD studies, but also from GWAS performed for many other complex traits over the last few years, that increasing sample size is indeed an effective approach in the hunt for genetic risk variants. For example, three GWASs from 2008 for adult human height with a combined sample of ~63,000 individuals reported a total of 54 SNPs[5], while in 2014, a GWAS including ~253,000 individuals identifying 697 significant variants[6]. Similar progress in identified risk loci is seen for schizophrenia. In 2009, the first genomic locus that was robustly associated with schizophrenia was discovered with a sample of ~3000 cases[7]. When in 2014 the sample size was increased to ~35,000 cases, 108 independent loci were discovered[8]. At present, the PGC data includes over 65,000 cases and almost 88,000 controls and the GWAS revealed 248 independent risk loci[9]. Furthermore, we recently extended our insomnia complaints GWAS described in chapter 6. We meta-analyzed the full genetic-data release of UK Biobank with 23andMe data, thereby increasing the sample size to the massive number of 1,331,010 participants, which increased the identified risk loci from a handful to no less than 202 independent loci[10]. Although each individual risk locus has a small effect on the variance in insomnia complaints, the large amount of loci made it feasible to inform us about causal mechanisms. These efforts of large-scale analyses will bring us closer in understanding the biology underlying complex traits, which I will discuss in one of the following sections of this chapter.

2.2 Gene-set analysis provides biological insight

In chapters 3 and 4, I applied gene-set analysis to investigate biological pathways that could contribute to the risk to develop psychiatric disorders. Gene-set analysis has two major advances compared to single SNP methods. The first one is a statistical advantage; evaluating the joint effect of multiple SNPs increases the statistical power to detect associations with the phenotype, as it reduces multiple testing. The second advance concerns the identification of biological mechanisms. Given the polygenic nature of many traits, it is likely that numerous SNPs of small effect aggregate in genes that share a similar cellular function. Identifying sets of genes with the same function will result in direct insight in the molecular or cellular mechanisms underlying the trait, while single SNP associations do not necessarily lead to this knowledge.

Gene-set analysis has been successfully applied for individual disorders and has led to novel biological hypotheses. For example the identification of neuronal calcium signaling in schizophrenia[8] and FMRP targets in autism spectrum disorder[11]. In chapter 4 we discovered gene sets that play a role across multiple psychiatric disorders, including the synaptic sets which corresponds with another cross-disorder pathway study that was performed on a smaller dataset[12]. At present, GWAS publications generally include gene-set analyses to aid the biological interpretation of GWAS results, likewise the GWASs of insomnia complaints

in chapter 6 of this thesis. In addition to the database pathways that are widely used, we extended the gene-set analysis in chapter 4 with tissue-specific gene expression profiles, which pointed to the involvement of brain tissues. Including gene expression data in gene-set analyses can provide additional insight. For example, we recently investigated tissue and single-cell expression data for insomnia complaints, which resulted in the identification of cortical structures and neuron cell types. In particular cell types from the claustrum, hypothalamus and striatum were associated, with the strongest association for medium spiny neurons. Together these results implicate involvement of a striato-cortical network, the regulation of reward processing, and the control of sleep and arousal in insomnia. These successes show the general ability of gene-set analyses to identify biological mechanisms underlying neuropsychiatric traits, which could reveal novel treatment possibilities, as these mechanisms make considerably larger and better drug targets than individual genes[13].

Nevertheless, the statistical power to detect gene-set associations can be limited by lack of genetic signal due to small sample size (see discussion above), which might be an explanation why we didn't identify synaptic gene sets for ADHD in chapter 2. The recent study of the PGC performing GWAS on the largest ADHD sample to date[4] included a gene-set analysis of all Gene Ontology (GO) gene sets, which did not yield significantly associated pathways, despite sufficient statistical power to identify the first 16 independent significant loci for ADHD, and a SNP heritability of 0.22. Yet synaptic gene sets were among the top gene sets, suggesting a potential role for synapse function in ADHD (of note, these GO gene sets were among the top sets of the cross-disorder gene-set analysis reported in chapter 4 as well, implying that involvement of these sets are not restricted to ADHD risk). In addition, a recent study reported the association of a glutamatergic gene set with severity of hyperactivity/impulsivity in an ADHD case-only sample[14]. A complication in gene-set analysis for highly heritable traits is that increasing sample size improves statistical power to only a limited extent[15]. This is because the associations of the genes in the gene set are compared to the associations of other genes in the genome. When many strong gene associations are present in the GWAS data – which is expected when a trait is highly polygenic – it is more difficult to distinguish the associations that are specific to the gene set from the many other associated genes.

Several tools for gene-set analysis are available. Although the underlying statistical structure is essentially the same, the tools show many differences. For instance, the type of null hypothesis often differs between methods, which affects how the results of the analysis should be interpreted. In chapter 3 we applied a competitive test in JAG to evaluate if a gene set is more strongly associated with the phenotype compared to other randomly generated gene sets, by using a permutation procedure to calculate empirical P values. However, we

did not correct the JAG analysis for gene size and gene density, which could have resulted in confounding effects by large genes, or multiple genes at the same locus that are within the same gene set. A recent review provided an extensive statistical evaluation of multiple gene-set analysis tools[15]. Biases due to factors such as gene size and linkage disequilibrium (LD) are problems that the majority of existing tools do not sufficiently address. Two tools properly correct for these possible biases and perform well on type-I error rate. The first test is the competitive test in MAGMA; a mean based method that tests whether the genes in a gene set are more strongly associated with the phenotype than the other genes in the genome. The second tool is INRICH; a count-based method that tests whether the proportion of P values below a given threshold in the set is greater than the proportion outside the set. However, setting the threshold is an arbitrary choice that can induce significant associations driven by one or a few genes rather than the biological function of the whole set. Hence, in chapter 4 we chose MAGMA to evaluate the importance of biological pathways for psychiatric disorders, as this tool utilizes all information available in the set of genes, which ultimately leads to a better representation of the biological pathway.

In addition to the tool that is applied, the results of gene-set analysis highly depend on the included gene sets. The majority of the gene-set studies have utilized publicly available databases. However, these gene sets are often incomplete and neither error-free nor unbiased, especially with regard to genes active in the brain[16,17]. Hence, we additionally investigated expert-curated gene sets in chapters 3 and 4. Although these sets can provide additional insight in causal mechanisms, the biological annotations of the current expert-curated gene sets may still not be optimal. Accurate annotation of genes to gene sets is a challenge, as current knowledge of gene functions is incomplete. Annotation should ideally be an ongoing process incorporating the constantly growing biological insight. Also, effects of regulatory variants outside the genes are currently not taken into account. Moreover, the observation that many genes practice more than one function or participate in multiple pathways makes annotation even more challenging. Overlap between gene sets might result in misleading conclusions when ignored. Conditional analyses can provide additional insight, likewise in chapter 4 where we reduced the significantly associated gene sets to a few independent sets. It is important to recognize the possible impact of these issues on gene-set analysis and pay attention to these in future studies. Still, future gene-set analyses will likely provide important progress in understanding the etiology of neuropsychiatric traits.

2.3 Widespread shared genetic effects on traits

The interest in studying the shared genetic effects on traits and diseases has grown over the last years, as it can inform about the causal relationships between traits and the molecular functions of genes[18]. One way to detect and quantify the level of pleiotropy, i.e. genetic

variants that affect multiple phenotypes, is by estimating the overall genetic correlation between traits using high-density SNP data for variance component analysis (GCTA[19] and BOLT-REML[20]) or polygenic risk score analysis[21,22]. These approaches consider the effects of all SNPs, including those that do not reach genome-wide significance, as heritability for complex traits is likely distributed over thousands of variants with small effects that have not been significantly associated by GWAS yet. Because individual-level genotype data is a requirement for these methods – and often this is not available due to informed consent constraints – a new approach was developed for estimating genetic correlations. This method, called cross-trait LD Score regression, only requires GWAS summary statistics[23]. This development resulted in a wide application of estimating genetic overlap, and it is now a typical analysis included in most GWAS publications, among those the insomnia complaints GWAS described in **chapter 6** of this thesis. We reported in these studies that insomnia complaints show positive genetic correlations with neuropsychiatric and metabolic traits, and negative correlations with subjective well-being and educational attainment. Furthermore, the results suggested that insomnia complaints more closely resemble neuropsychiatric traits than other sleep-related traits. These observations give insight in shared etiology and provide a starting point for further studies of these relationships.

Despite the usefulness of estimating genetic correlations, the parameter of genetic correlation only quantifies the overall proportion of the genetic loci that are shared by two traits; it does not inform us which loci are shared or which biological pathways are involved. In **chapter 4** we aimed to increase our insight in the biological pathways that actually account for the genetic overlap between psychiatric disorders. Overlap between disorders was already shown at the SNP-level in earlier studies[24,25]. A specific example is *CACNA1C* that has been associated with bipolar disorder, schizophrenia and major depressive disorder[26]. In addition, several cross-disorder pathways have been identified indicating that genetic overlap is nonrandom at the molecular or pathway level[12]. We chose the gene-set analysis approach to capitalize on recent increased genetic signal in GWASs of psychiatric disorders and further investigated the genetic overlap between five disorders. We identified sets of genes that are related to the postsynapse, neuron projection and expression profiles of frontal cortex and cerebellar hemispheres, suggesting a common role of these cellular components and brain areas in general vulnerability for multiple psychiatric disorders.

When studying pleiotropy, one should consider that pleiotropy might be induced by confounding causation. In **chapter 6** we identified *MEIS1* as the strongest genetic signal for insomnia complaints, but this gene has robustly been related to RLS previously, a syndrome that is symptomatically related to insomnia. In order to draw conclusions, we investigated a possible confounding effect of RLS using an additional sample with RLS patients and

multiple analysis techniques. We found evidence for an independent effect of *MEIS1* on insomnia, although RLS likely influence the association signal as well. It is challenging to infer causal effects from statistical analyses of genetic data only, as unobserved phenotypes that are genetically correlated with the investigated phenotype may be the actual causal associaton[18]. Extended phenotyping of cohorts will be required to investigate the role of causation in pleiotropy in more detail.

Besides the genetic overlap between neuropsychiatric traits, evidence for widespread pleiotropy has been reported for other traits as well[18,23,27]. To interpret these findings we should consider that causal variants are surprisingly uniformly distributed over the genome[28]. It has been estimated that 71% to 100% of all 1-Mb windows in the genome include at least one risk variant for schizophrenia[20]. The observation that the thousands of associated genetic variants for complex traits are spread broadly across the genome implies that pleiotropy may be ubiquitous[29]. The true nature of the detected pleiotropy is currently unknown. Methods like Mendelian Randomization generally assume that pleiotropy is mainly induced by traits that are causally related[30]. However, another explanation might be that a genetic variant affects multiple traits not because the traits are directly causally related (i.e. the traits are influenced by genes with direct relevance to the traits), but because the same cell type(s) are involved and hence the traits are regulated through genes that are part of the same cell-specific regulatory network[31]. The recently proposed 'omnigenetics model' predicts that variants with regulatory effects in a specific tissue are likely to have effects on all traits that are influenced through that tissue31. Future studies investigating these possibilities will be essential to understand the biology underlying pleiotropy.

2.4 Contribution of rare variants

GWASs resulted in an enormous increase in the number of known genetic loci for complex traits. Nevertheless, GWAS investigates the contribution of common variants only, and other types of genetic variation including rare variants, structural variants, and epigenetic variation likely make up part of the genetic architectures of complex traits as well. Concerning rare variants, it has been estimated that they contribute half the heritability that is explained by common variants[32]. Studying the role of rare variants in complex traits is still in its infancy, but the emergence of exome-wide and genome-wide sequencing provides the first insights. For multiple traits, an inverse relationship has been found between the effect size and the minor allele frequency (MAF) in the population[33]. This observation could be explained by selective pressure on genetic variants that strongly contribute to the risk of developing traits that affect fitness.

The first association studies of rare variants suggested that some complex disorders like ASD are skewed towards rarer risk variants compared to other complex disorders and diseases like

schizophrenia, type II diabetes, and cardiometabolic traits[34]. Concerning ASD, early findings already pointed to an import role of rare variants (including *de novo* mutations and copy number variations), with convergence of results across large samples and laboratories[35]. These findings account for the enormous progress in ASD gene discoveries during the last decade, and offers an important opportunity to identify the causal neurodevelopmental mechanisms which might aid in treatment development[36]. The identified ASD genes point most consistently to involvement of synaptic functions, but also a wide range of other cellular functions, including chromatin modification, receptor signaling, and protein synthesis[35]. Although less apparent than for ASD, rare variants have been identified for schizophrenia as well. These variants implicate proteins involved in plasticity at glutamatergic synapses, especially the ARC and NMDAR complexes, FMRP targets, and voltage-gated calcium ion channels[37–40]; mechanisms that have been highlighted by common variant studies as well, including chapter 4 of this thesis.

Although sequencing costs decrease and the quality increases by higher sequencing throughput, it is still costly to assess the large sample sizes that are required for studying complex traits. The ExomeChip has been developed to offer an interim approach. This genotyping array allows studying rare variation within the exome in a cost-effective way for large sample sizes. Hence, we applied this strategy in chapter 5 to investigate the genetic architecture of substance use. Despite the large sample size that we achieved by meta-analyzing data from multiple cohorts, we could not identify rare variants for alcohol consumption and tobacco use. Power analysis showed that with this large sample size we would be able to detect rare variants that have substantial effects on the phenotypes. However, large effect sizes do not have to be a general property of rare variants, and it can well be that rare variants with small effects do play a role. Interestingly, a recent ExomeChip study on smoking and drinking behavior with sample sizes a magnitude higher than our study, ranging from 70,847 to 164,142 individuals, did still not provide robust evidence for associated rare variants[41]. Larger sample sizes will be required to investigate if rare variants with effect sizes similar to those found for common variants contribute to the genetic architecture of addiction behaviors.

Apart from the possibility of underpowered studies, other factors could explain the lack of significant associations in most Exomechip studies thus far[34]. A limitation of the ExomeChip is that it only genotypes ~250,000 rare and low-frequency variants. It is likely that many additional functional exonic variants exist that are currently missing on the chip. Furthermore, rare variants outside the exome will be missed. Although an enrichment analysis of functional and regulatory features of rare variants using UK10K data showed that exonic variants represent the strongest degree of enrichment in trait associations, non-coding low frequency variants contribute as well[42].

Even though we currently do not know how much of the total additive genetic variation of complex traits can be attributed to rare and low-frequency variants, it has been estimated from imputed genotype data for height that more additive genetic variation is explained by these variants than expected under an evolutionary neutral model, which is consistent with purifying selection of the height-associated variants[43]. Population-scale whole-genome sequencing is increasingly used to systematically study the associations of rare variants and their properties. In the near future, these data will be available in a fast growing number of samples, and the contribution of rare and low-frequency variants to complex traits can then be estimated more precisely.

2.5 From genetic variants to biological pathways – Are we there yet?

The number of identified genetic loci for neuropsychiatric and other complex traits rapidly increased the last decade. These exciting novel findings are largely driven by the exponentially increase in sample sizes, which I discussed in the first part of this discussion. While this step was critical for the field to move forward, the next challenge is to annotate, prioritize and interpret this wealth of GWAS output data. This path from GWAS to biology is not straightforward, as an association between a genetic variant and a trait is not directly informative. First of all because an associated variant is not individually associated, but it is part of a genomic region including many correlated SNPs. In turn, these SNPs may cover multiple genes of which some may be relevant to the trait, while others are not. Based on the GWAS association results alone we cannot identify the true causal genes. A second complication is that the majority of GWAS hits are located in non-coding or intergenic regions[44], making inference of the underlying causative genes difficult.

Recently, a range of secondary analysis tools has been developed to understand the biological implications of GWAS results. One strategy is to aggregate all associated variants to identify the biological pathways and cellular mechanisms that drive the trait. Gene-set analysis is such a method, which I extensively discussed in one of the previous sections. Its use has been proven by the widespread successful application using sets of genes based on shared biological pathways or cellular functions (see e.g. Sullivan and Posthuma, 2015[45] for an overview of the gene-set findings for psychiatric disorders). Future application of gene-set analysis with extensions of new data types like the tissue-type expression data we included in **chapter 4**, but also single-cell expression data and data on regulatory functions and methylation data, is expected to result in new insights in the biological underpinnings of neuropsychiatric traits.

A second strategy is to dissect the functional impacts of individual associated variants. The development of molecular technologies and new analytical methods offer opportunities

to increase our understanding of the biological implications of the identified genetic loci, especially for the non-coding variants. For example, information on open or closed chromatin (i.e. the DNA is wrapped around a complex of proteins in a state that the DNA is open or closed to entry of the molecular machinery required for transcription) for different tissue types revealed that GWAS-associated variants are highly enriched for regions of active chromatin such as promotors and enhancers; an observation that was particularly seen in the tissue types relevant to the studied traits[31]. The projects GTEx[46], ENCODE[47] and Epigenome RoadMap[48] contribute essential maps of regulatory annotation to interpret the roles of non-coding variants. GTEx provides for many tissue types data on expression quantitative trait loci (eQTL), the influence of genetic variants that affect the expression of genes. ENCODE and Epigenome RoadMap integrated over a dozen types of experimental data that give information on histone modification patterns, DNA accessibility, DNA methylation and gene expression in multiple tissue types, and established maps of regulatory elements and activators and repressors of regulatory modules. The application of this intriguing and insightful data to the interpretation of GWAS results will likely become increasingly apparent, as the projects move forward quickly and provide additional data and insights. For example, very recently the GTEx consortium presented the second phase of their project, in which they report that the expression levels of almost all genes in the genome are affected by genetic variation[49]. Most of these genetic variants are located in the proximity of the affected gene (*cis*-eQTLs), but moreover several hundred genetic variants were identified that act over long distance (*trans*-eQTLs). Interestingly, *cis*-eQTLs tend to alter gene expression on most tissue types, while *trans*-eQTLs tend to affect expression in only one or a few tissues. In addition, the consortium addressed for the first time on a genome-wide scale the relevance of rare variation in determining gene expression[50]. These GTEx maps open the possibility to study possible regulatory effects in different tissue types of non-coding common variants identified by GWAS, and rare non-coding variants in forthcoming whole-genome sequencing studies. Furthermore, specifically for the neuropsychiatric research field, the recently initiated projects CommonMind and PsychENCODE[51] will provide valuable data on regulatory elements, epigenetic modifications, and gene and protein expression in tissue- and cell-type specific samples from post-mortem brains of individuals with neuropsychiatric disorders. These exciting maps of regulatory functions will aid the interpretation of GWAS results of neuropsychiatric disorders and traits in future studies.

Especially the integration of these different types of data can guide in pinpointing the most likely variants and genes in associated genomic loci, as multiple lines of evidence will provide more robust conclusions. However, extracting all these data and merging them is challenging. To provide a systematic screening of multiple databases and combining and prioritizing information of interest, several tools have been developed, e.g. PrixFixe[52] and FUMA[53].

PrixFixe prioritizes genes based on co-functional networks that are created based on genetic, protein and transcription binding interactions, and shared GO terms and expression patterns. FUMA functionally annotates identified GWAS loci and prioritizes the most likely causal SNPs and genes utilizing information from 14 state of the art biological databases and tools, including functional consequences on gene functions, deleteriousness, regulatory functions, gene expression of various tissue types, biological pathways and disease and drug target information. These extensive annotations facilitate the generation of hypotheses for causal variants and genes which could be used for functional follow-up experiments.

Follow-up functional studies *in vitro* or *in vivo* are important to study the actual biological consequence of detected variants. One of the successes was the identification the complement component 4 (C4) genes as the causal genes in the major histocompatibility complex (MHC) that is the strongest genetic association for schizophrenia[54]. Multiple experiments showed that many variants influenced *C4* gene expression levels in the brain, and that *C4* genes mediate synapse elimination. Despite successes of follow-up studies, performing similar extensive functional experiments on the wealth of identified loci will be time consuming, even if integrative annotation tools like FUMA have prioritized genes. Because genetic variants likely act jointly through complex cellular networks, there is a great need to develop high-throughput cell-based model systems that can investigate networks of genes in different cell types. First efforts of automated high-throughput functional screening emerge, such as CRISPR[55] technology and induced pluripotent stem cells (iPSCs)[56], and will be promising methods to elucidate causal mechanisms in future.

Especially iPSC technology provides a promising technique to study polygenic traits like neuropsychiatric disorders[57]. iPSCs are generated from patient-derived cells, primarily fibroblasts or blood cells. Patients can be selected based on a high genetic risk for a disorder due to the accumulation of many common variants with small effects or a rare variant with large effect[58]. Not only these known variants will be under investigation, but the complete genetic background of an individual will be captured, including possible genetic factors that are not yet identified[57]. The advent of iPSC research is that the cells can be programmed to become every other cell type in the human body (e.g. glutamatergic neurons and astrocytes) on which a range of methods of cell phenotyping can be applied, including transcriptomics, proteomics, cell morphology, functional activity and connectomics[58]. These experimental methods can point to potential pathological mechanisms of neuropsychiatric disorders. Moreover, the development of the cells can be followed, allowing to study the progress of disease processes. The first successes have been made in iPSC research for neuropsychiatric disorders. For example one of the first iPSC studies of schizophrenia reported that iPSC-derived neurons showed decreased connectivity, synapses, spine density

and expression of glutamate receptors[59]. Treatment of the iPSC-derived neurons with the dopaminergic antagonist loxapine, but not clozapine, olanzapine, risperidone or thioridazine, during the last three weeks of neuronal differentiation increased neuronal connectivity in all patient lines. This finding illustrates the additional value of iPSC studies for *in vitro* treatment screening. Although many challenges remain for this system, iPSCs will offer a powerful platform for drug discovery and personalized medicine in the near future[56].

3. Final conclusions

Major advances have occurred in the genetics of neuropsychiatric traits over the recent years, in which the exponential increases in sample sizes played an instrumental role. Efforts are being made to maintain this progress in sample sizes over the next years, which will result in additional insight[60]. Current GWAS results have shown that most neuropsychiatric traits are highly polygenic, mostly influenced by large numbers of weakly associated variants. Although the identification of genetic loci is crucial for the field to move forward, so too is the development of tools to biologically interpret this wealth of data. Several methods have been developed that place GWAS results in a broader biological context, and first biological insights emerge. Nevertheless, many challenges remain to fully understand the complexity of the etiology of neuropsychiatric traits. The current achievements of identifying hundreds of loci for complex traits show that the field is ripe for further exploration in the light of novel types of data. Insights are expected to grow with new datasets of expression levels and regulatory functions in neuropsychiatric disorders (CommonMind, PsychENCODE) that will be available in the very near future.

In addition, large-scale whole-genome sequencing is a promising next step to estimate explicitly the contribution of low-frequency, rare and structural variants to the genetic architecture of neuropsychiatric traits. It is expected that this will lead to discoveries of highly informative rare alleles that are associated with diseases. These variants could contribute to the validation of therapeutic targets, for example variants could mimic the effect of modulating drug target genes, thereby informing if the drug could be successful[34].

In this thesis I have shown that neuropsychiatric genetics is a fast evolving field, and I expect that our understanding of the etiology of neuropsychiatric traits will quickly evolve during the coming years. One of the ultimate objectives of genetic research for neuropsychiatric traits is to inform the development of new therapeutic targets. The diverse routes of genetic research will drive future progression in the development of more effective treatment and possibly even prevention of neuropsychiatric disorders.

References

1. Groen-Blokhuis, M. M. *et al.* Attention-deficit/hyperactivity disorder polygenic risk scores predict attention problems in a population-based sample of children. *J. Am. Acad. Child Adolesc. Psychiatry* **53,** 1123–1129.e6 (2014).
2. Martin, J., Hamshere, M. L., Stergiakouli, E., O'Donovan, M. C. & Thapar, A. Genetic risk for attention-deficit/hyperactivity disorder contributes to neurodevelopmental traits in the general population. *Biol. Psychiatry* **76,** 664–671 (2014).
3. Stergiakouli, E. *et al.* Shared genetic influences between attention-deficit/hyperactivity disorder (ADHD) traits in children and clinical ADHD. *J. Am. Acad. Child Adolesc. Psychiatry* **54,** 322–327 (2015).
4. Demontis, D. *et al.* Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv* (2017). doi:https://doi.org/10.1101/145581
5. Visscher, P. M. Sizing up human height variation. *Nat. Genet.* **40,** 489–490 (2008).
6. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
7. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
8. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
9. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Presented at 25th World Congress of Psychiatric Genetics, Orlando, Florida, USA. (2017).
10. Jansen, P. R. *et al.* Genome-wide study of insomnia (N=1,331,010) identifies novel genetic loci and functional pathways. *bioRxiv* (2018). doi:https://doi.org/10.1101/214973.
11. Jansen, A. *et al.* Gene-set analysis shows association between FMRP targets and autism spectrum disorder. *Eur J Hum Genet* **25,** 1–6 (2017).
12. The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18,** 199–209 (2015).
13. Yauch, R. L. & Settleman, J. Recent advances in pathway-targeted cancer drug therapies emerging from cancer genome analysis. *Curr. Opin. Genet. Dev.* **22,** 45–49 (2012).
14. Naaijen, J. *et al.* Glutamatergic and GABAergic gene sets in attention-deficit/hyperactivity disorder: association to overlapping traits in ADHD and autism. *Transl. Psychiatry* **7,** e999 (2017).
15. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. Current issues in gene-set analysis: statistical properties. *Prep.*
16. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4323–4328 (2008).
17. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
18. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48,** 709–717 (2016).
19. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
20. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47,** 1385–1392 (2015).
21. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
22. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (2013).
23. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47,** 1236–1241 (2015).

24. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
25. 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* **381,** 1371–1379 (2013).
26. Green, E. K. *et al.* The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol. Psychiatry* **15,** 1016–1022 (2010).
27. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89,** 607–618 (2011).
28. Yang, J. *et al.* Genome-partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
29. Visscher, P. M. & Yang, J. A plethora of pleiotropy across complex traits. *Nat Genet* **48,** 707–708 (2016).
30. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12,** 204–213 (2011).
31. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
32. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49,** 986–992 (2017).
33. Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 18026–18031 (2011).
34. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18,** 77 (2017).
35. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet. Neurol.* **14,** 1109–1120 (2015).
36. Gamsiz, E. D. *et al.* Discovery of Rare Mutations in Autism: Elucidating Neurodevelopmental Mechanisms. *Neurotherapeutics* **12,** 553–571 (2015).
37. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
38. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–190 (2014).
39. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17,** 142–153 (2012).
40. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146,** 247–261 (2011).
41. Liu, D. J. *et al.* Exome chip meta-analysis elucidates the genetic architecture of rare coding variants in smoking and drinking behavior. *bioRxiv* (2017). doi:https://doi.org/10.1101/187658
42. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
43. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47,** 1114–1120 (2015).
44. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).
45. Sullivan, P. F. & Posthuma, D. Biological pathways and networks implicated in psychiatric disorders. *Curr. Opin. Behav. Sci.* **2,** 58–68 (2015).
46. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
47. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
48. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
49. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
50. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550,** 239–243 (2017).

51. Akbarian, S. *et al.* The PsychENCODE project. *Nat Neurosci* **18,** 1707–1712 (2015).
52. Tasan, M. *et al.* Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat. Methods* **12,** 154–159 (2015).
53. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. FUMA: Functional mapping and annotation of genetic associations. bioRxiv (2017). doi:https://doi.org/10.1101/110023
54. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530,** 177–183 (2016).
55. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157,** 1262–1278 (2014).
56. Soliman, M. A., Aboharb, F., Zeltner, N. & Studer, L. Pluripotent stem cells in neuropsychiatric disorders. *Mol Psychiatry* **22,** 1241–1249 (2017).
57. Marchetto, M. C., Brennand, K. J., Boyer, L. F. & Gage, F. H. Induced pluripotent stem cells (iPSCs) and neurological disease modeling: Progress and promises. *Hum. Mol. Genet.* **20,** 109–115 (2011).
58. Falk, A. *et al.* Modeling psychiatric disorders: from genomic findings to cellular phenotypes. *Mol. Psychiatry* **21,** 1167–1179 (2016).
59. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473,** 221–225 (2011).
60. Sullivan, P. F. *et al.* Psychiatric Genomics: An Update and an Agenda. *bioRxiv* 115600 (2017). doi:10.1101/115600