Genetic architecture of childhood neuropsychiatric and involuntary movement disorders

Muhammad Sulaman Nawaz

Thesis for the degree of Philosophiae Doctor

December 2022

School of Health Sciences

FACULTY OF MEDICINE

UNIVERSITY OF ICELAND

Genetic architecture of childhood neuropsychiatric and involuntary movement disorders

Muhammad Sulaman Nawaz

Thesis for the degree of Philosophiae Doctor

Supervisor(s)

Kari Stefansson, Professor, University of Iceland,
Founder and CEO, deCODE Genetics, Iceland

Tutor (**Supervisory teacher**)

Hreinn Stefansson, Head of CNS Division, deCODE Genetics, Iceland

Doctoral committee

Þorgeir Þorgeirsson, Director of Genetics, deCODE genetics, Iceland
Gisli Masson, VP of Informatics, deCODE Genetics, Iceland
Guðbjörn F. Jonsson, Bioinformatician, deCODE Genetics, Iceland
Petur Luðvigsson, Clinical Associate Professor, Barnaspitali Hringsins, Iceland

December 2022

School of Health Sciences

FACULTY OF MEDICINE

UNIVERSITY OF ICELAND

Erfðafræði þroskaraskana, ósjálfráðrar hreyfingar, hvatvísi og þráhyggju

Muhammad Sulaman Nawaz

Ritgerð til doktorsgráðu

Leiðbeinandi/leiðbeinendur

Kari Stefansson, Professor, University of Iceland,
Founder and CEO, deCODE Genetics, Iceland

Doktorsnefnd

Hreinn Stefansson, Head of CNS Division, deCODE Genetics, Iceland
Þorgeir Þorgeirsson, Director of Genetics, deCODE genetics, Iceland
Gisli Masson, VP of Informatics, deCODE Genetics, Iceland
Guðbjörn F. Jonsson, Bioinformatician, deCODE Genetics, Iceland
Petur Luðvigsson, Clinical Associate Professor, Barnaspitali Hringsins, Iceland

Desember 2022

Heilbrigðisvísindasvið
LÆKNADEILD
HÁSKÓLI ÍSLANDS

ORCID: https://orcid.org/0000-0002-5576-9007

Reykjavik, Iceland 2022

Ágrip

Taugaþroskaraskanir á hvatvísis-þráhyggju rófinu eru þrálát og hamlandi einkenni sem koma oft fram snemma á lífsleiðinni. Há tíðni fylgiraskana finnast samhliða Tourette heilkenni (TS), kækjum (Tics), þráhyggju- og árättu- hegðun (OCD), athyglisbrest með og án ofvirkni röskun (ADHD), og einhverfurófsröskun (ASD) og dæmi eru um að sömu algengu breytileikarnir auka áhættu á mismunandi röskunum.

Þessi ritgerð fjallar um erfðafræði rúmmáls heila og fimm taugaþroskaraskana á hvatvísis-þráhyggju rófinu (Tourette, Tics, OCD, ADHD, og fótaóeirð (RLS)) með það að markmiði að finna nýja breytileika sem ávísa áhættu af því að þróa þessar svipgerðir og að rannsaka áhættu á krossröskun. Erfðafræði þessara raskana er flókin. Þrátt fyrir stórar safngreiningar (meta-analysis), hafa tiltölulega fáir breytileikar fundist sem ávísa áhættu á þessum röskunum, né varpað ljósi á þá líffræðilegu ferla sem þar liggja að baki. Betri skilningur á erfðafræðilegri undirstöðu þessara raskana gæti stórbætt og flýtt fyrir greiningu, gefið betri innsýn inn í sjúkdómsferlið og bent á nýjar lyfjameðferðir.

Eintakabreytileikar annaðhvort koma fyrir innskotum eða eyða út svæðum í erfðamenginu og hafa þannig áhrif á hversu mikið er tjáð af genum sem eru á þeirra áhrifasvæði. Leit í líklegum genum staðfestir að úrfelling á *AADAC* geninu er áhættu þáttur fyrir Tourette heilkenni. Þar að auki, höfum við sýnt fram á tengls á milli ákveðinna eintakabreytileika, sem hafa áður verið tengdir við einhverfu og geðklofa, og ADHD. Þessi fylgni varpar ljósi á þá erfðafræðilegu áhættu sem ADHD deilir með einhverfu og geðklofa. Safngreining niðurstaðna úr víðtækri erfðamengisleit skilaði engum erfðabreytileikum sem voru í marktægt hærri tíðni í TS en í viðmiðunarhópi. Þó staðfestir samanlagt arfgerðar-skor fjölgena eðli Tourette heilkennis.

Stórar safngreiningar hafa auðkennt nýja breytileika sem ávísa áhættu á röskun á hvatvísis-þráhyggju rófinu. Staðsetning breytileika og aðrar aðferðir hafa verið notaðar til að tengja breytileika við gen og Mendelsku slembivali beitt til að rýna í orsakasamhengi, með fókus á líffræðilega ferla þessi gen taka þátt í.

Breytileiki í rúmmáli heilans getur haft áhrif á tengsl á milli uppbyggingu og virkni í heilanum. Stór safngreining niðurstaðna úr víðtækri erfðamengisleit á rúmmáli heilans hefur leitt í ljós 64 breytileika í erfðamenginu sem útskýra um

5.0% dreifni svipgerðarinnar. Erfðafylgnigreining (GC) sýnir jákvæða fylgni á milli rúmmáls heilans og vitrænna hæfileika og taugasjúkdóma.

Breið svipgerða erfðafylgni greining og orsakagreiningar voru notaðar til að kryfja flókin fjölgena sambönd á milli TS, kækja, ADHD, OCD og RLS. GC þessara fimm raskana á móti 1,140 birtra niðurstaðana úr heilgenóms erfðatengslaleitum, bar kennsl á 59 algeng svipbrigði sem eru sýna marktæka erfðafylgni (FDR < 0.05). Stigveldis þyrpingagreining á þessu 59 svipgerðum leiddi í ljós fimm dulda klasa; (1) taugaþroska- og geðraskanir, (2) tilfinningar raskanir, (3) útlima og vöðva verkir, (4) offita / óheilbrigður lífstíll, (5) vitsmuna / lærdóms svipgerðir.

Sjaldgæfir og algengir breytielikar hafa verið tengdir við TS, ADHD, RLS og rúmmál heilans. Lykil spurning er hvort breytileikarnir, sem hafa verið tengdir við breytingar í heila, valdi taugasjókdómum í gegnum áhrifin sem þeir hafa á heilann, eða hvort erfðafræðileg tilhneiging til að þróa taugasjókdóma hefur áhrif á uppbyggingu og þroska heilans. Tvíátta Mendelsk slmbivals greining á 34 svipgerðum, með innbyrgðis fylgni, samanborið við rúmmál heilans gaf til kynna að rúmmál heilans annað hvort hefur áhrif á taugaþroska raskanir (ADHD) og taugasjókdóma (Parkinson's) eða þá að orsakasamhengið sé stýrt af svipgerðum sem sýna sterka fylgni við rúmmál heilans.

Lykilorð:

Erfðafræði, víðtæk erfðamengisleit, Tourette heilkenni, Athyglisbrestur með án ofvirkni, Áráttu- og þráhyggjuröskun, fótaóeirð, heila rúmmál.

Abstract

Neurodevelopmental disorders on the impulsivity-compulsivity spectrum are chronic disabling conditions with an early onset. High rates of comorbidity have been reported between Tourette syndrome (TS), Tics disorder (Tics), obsessive-compulsive disorder (OCD), attention-deficit/hyperactivity disorder (ADHD) and autism spectrum disorder (ASD) and these disorders share cross-disorder risk, conferred by common variants.

In this thesis the focus is on the genetics of human intracranial volume (ICV) and five impulsivity-compulsivity spectrum neurological disorders (Tourette, Tics, OCD, ADHD, and restless legs syndrome (RLS)) with the aim of finding novel sequence variants conferring risk of these behaviours and study their cross-disorder risk. The genetics of these disorders is complex. Despite large meta-analyses, relatively few sequence variants have been associated with these disorders and perturbed biochemical pathways have not been clearly outlined. Better understanding of their genetic underpinnings may greatly improve and accelerate diagnosis, give insights into disease processes, and point to novel targets for drug therapies.

Copy number variations (CNVs) introduce insertion/deletion throughout the genome thereby impacting gene expression through gene dosage effect. A candidate gene study confirms *AADAC* deletion as a risk factor for TS. Moreover, a group of rare, recurrent CNVs, so called neuropsychiatric CNVs, confer high risk of ADHD. This association highlighted shared genetic risk of ADHD with ASD and schizophrenia. The GWAS meta-analysis of TS didn't find any significant association while aggregate risk score of common variants confirmed polygenic nature of TS.

Through large meta-analyses, more sequence variants conferring risk of diseases of the impulsivity-compulsivity spectrum have been uncovered. Colocalization analyses were used to identify affected genes and Mendelian randomization to search for causal relationships, with a focus on the biological insights these associations are beginning to produce.

Variations in ICV can impact brain structure-function relationships. The GWAS meta-analysis of ICV uncovered 64 sequence variants explaining 5.0% of the

trait's variance. Genetic correlation analysis shows positive correlation between ICV and cognitive abilities and neurological traits.

Phenome-wide genetic correlation (GC) and causal analyses were used to dissect complex polygenic nature of TS, Tics, ADHD, OCD, and RLS. GC of these five disorders with 1,140 published GWAS studies identified 59 common genetically correlated traits (FDR < 0.05). The hierarchical clustering of 59 correlated traits identified five latent clusters: (1) neuropsychiatric or neurotic disorders, (2) emotional disorders, (3) peripheral and muscular pain (4) obesity / poor lifestyle, and (5) cognition / learning traits.

Rare and common variants associate with TS, ADHD, RLS, and ICV. The key question is whether variants, associated with structural changes in the brain, cause neurological disorders through their effect on brain structure or alternatively whether genetic predisposition to certain neurological disorders impacts brain structure or development. Bidirectional MR analyses of 34 correlated disorders compared with ICV revealed that ICV either has a causal effect on a neurodevelopmental disorder (ADHD) as well as on a neurodegenerative disorder (Parkinson's) or these causal relationships may be driven by traits closely correlated with ICV.

Keywords:

Genetics, GWAS, Tourette syndrome, Attention deficit / hyperactivity disorder, obsessive compulsive disorder, restless leg syndromes, intracranial volume.

Acknowledgements

During the years that I have been doing my PhD, I have become greatly indebted to my excellent supervisors for making this journey an interesting and unforgettable one. Firstly, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Kari Stefansson, for providing me with opportunities to perform PhD research and to grow professionally. Prof. Kari's dedication, encouragement to execute challenging research tasks, his sincerity, and guidance I will never forget. He is ever inspiring to me for being a devoted scientific leader. His persona and professional achievements have been source of encouragement and ideas as I hurdled through the path of this PhD thesis, and undoubtedly a true definition of a scientific leader and the ultimate role model.

This thesis would not have been possible without my supervisory teacher, Dr. Hreinn Stefansson, whose guidance right from the 1st step in PhD research project enabled me to develop an understanding of the subject and professional working at deCODE Genetics, Iceland. I am thankful for the extraordinary experiences he arranged for me and for providing me with opportunities to grow professionally. His exemplary role of simplifying the research questions and sharing scientific story telling helped trained at presentation and communications. He always provided positive criticism, showed selflessness, modesty, and guided me how to exercise challenging task. It has been an honour to learn from Prof. Kari and Dr. Hreinn.

Words cannot express my gratitude to thank Dr. Þorgeir Þorgeirsson for his in-depth analytical skills, his mastery of presenting research work. I learnt from him to build up research work by connecting missing dots, how to be good at self-critique and still not overdo to step into imposter syndrome boundary. I am thankful for his great contributions which helped me furnish this thesis.

Special thanks to Dr. Gisli Masson for helping me with advanced Bioinformatics skills sets and modern IT network to perform statistical genetics at deCODE Genetics, Iceland. I appreciate his welcoming and pragmatic approach in dealing with research requests. I am also grateful to Dr. Guðbjörn F. Jonsson for arranging multiple practical sessions of Bioinformatics algorithms and software usage which helped performing research work. I would like to extend my sincere

thanks to Prof. Dr. Petur Luðvigsson for providing guidance to understand and assistance in arranging phenotype data of neurodevelopmental disorders. This work wouldn't have been possible without availability of phenotype data. Major part of this thesis is based on using advanced statistical approaches (developed or employed existing) to understand the genetic relationship of studies traits. All this work would not have been possible without the teaching, guidance, and supervison by Dr. Daníel F. Guðbjartsson, and Dr. Guðmar Thorleifsson. I owe a big thank and applause to both for their consistent guidance, discussions, and follow up of research work. Moreover, I would like to thank Dr. Patrick Sulem for helping me understanding and establishing methods for colocalization and monogenic diseases analysis.

I also could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise. Additionally, this endeavour would not have been possible without the generous support from the Marie Curie Initial Training Networks TS-EUROTRAIN (FP7-PEOPLE [Grant No. GA 316978]), and deCODE Genetics/Amgen, who financed my research. This endeavour would not have been possible without professional opportunity, guidance by colleagues, and accessibility of research data from deCODE Genetics/Amgen.

Countless people supported during my PhD research journey. I would to like especially acknowledge and thank for the wonderful experience of learning from /working with/collaborating with; G. Bragi Walters, Andres Ingason, Daníel F. Guðbjartsson, Gyða Bjornsdottir, Guðmundur Einarsson, Magnus O. Ulfarsson, Rosa S. Gisladottir, Larus J. Guðmundsson, Lina Jonsson, Astros T. Skuladottir, Unnur Unnsteindottir, Guðrun A. Jonsdottir, Maria Didriksen, Dongmei Yu, Sigurður H. Magnússon, Omar Gustafsson, Benedikt A. Jonsson, Mariana Bustamante, Gisli Halldorson, Sigurjón A. Guðjónsson, Anna Helgadottir, Guðmar Thorleifsson, Patrick Sulem, all co-authors, colleagues, and collaborators. All these colleagues/collaborators helped me find my footing as I started this process. Friends have priceless place in our lifes, we turn to them in best or worst times, they help us attain peace and joy in all situations. I owe a big thank to all of my friends for their professional and personal assistance, among those I would especially name Umair Seemab, Qaisar Farid, Qurrat ul Ain, Syed Shujaat Ali Zaidi, Hassan Khalid, Qaisar Fatmi, Muhammad Imran, Muhammad Waseem, Ammad ud Din, Wilayat Hussain, Hina, Sameera, Samra, Zahra, Rafia, Arshan, Hassan, Afraz Raja, Aziz ul Haq, Javed, Ihtisham, Taha, Wajid, Atef, and Waqar to name a few, thank you all.

My family deserves endless gratitude: my late father (Malik Muhammad Nawaz) for teaching me humbleness, love for knowledge, science, and nature. My mother, Salma Nawaz, for teaching me how to deal with difficult tides by keeping the self-esteem. My loving and selfless brothers (Atif, Kashif, Asim, Aamir, and Shoaib) thank you all for your unconditional love, support (moral, spiritual, and wordly). I will forever be indebted for such a care. Also thank you for teaching me that an assertion of dominance is not necessarily a bad thing. Special thanks and gratitude to eldest brother, Atif. Imran Malik, who has been more than a father to me. He is the one who introduced me to Bioinformatics and modern genetics when I was deserted by marginally missing medical school examination. My loving sister, Qurat ul Ain, who always encouraged me to take up challenges and continue until you succeed, thank you Anee and Nawaz family for your love and support that kept me motivated and confident.

Finally, I owe my deepest gratitude to Zahida, who is the love of my life and privilidge to call my better half. I am forever thankful for the unconditional love and support throughout the entire thesis process and beyond. My accomplishments and success are because you believed in me. Deepest thanks to our children (Yahya and Mirha), who keep me grounded, remind me of what is important in life, and are always supportive of dad's adventures. There have been countless evenings that they missed play time as I was working on thesis project. Yahya and Mirha, I am deeply indebted to you all for your love and support.

Lastly, I would be remiss of not mentioning my schoolteachers and undergraduate study professors, especially Mr. Tanveer Hussain, Mr. Mazhar Iqbal, Mr. Anwar, Mr. Adil Zubair, Mr. Raffaqat Hussain (late), Mr. Shafi, Mr. Raja, Ch. Lateef sb (late), Mr. Sagheer sb, Dr. Sajid Rashid, Dr. Amir Ali Abbasi, Dr. Sahar Fazl, Dr. Ayesha Fatima, and Mrs. Qurrat-ul-Ain. Their belief in me has kept my spirit and motivation high during this process.

Dedication

Heartly dedicates to my loving and selfless father, Malik Muhammad Nawaz, (late, الله رحمة), my courageous mother, Salma Nawaz, loving wife, Dr. Zahida Parveen, adoreable children, Yahya Sulaman Nawaz and Mirha Sulaman Nawaz, dearest siblings (Atif, Kashif, Asim, Aamir, Shoiab, and lovely Qurat ul Ain), and Nawaz family for their love, trust, and support.

Contents

List of abbreviations

ADHD: Attention deficit hyperactive disorder

ASD: Autism spectrum disorder

CNV: Copy number variation

DSM: Diagnostic and Statistical Manual

EAF: Effect allele frequency

EGGC: Early Growth Genetics Consortium

eQTL: Expression quantitative trait loci

GC: Genetic correlation

GWAS: Genome wide association scan

HC: Head circumference

ICD: International classification of diseases

ICV: Intracranial volume

IVW: Inverse-variance weighted

IVs: instrumental variables

LD: Linkage disequilibrium

LDSC: Linkage disequilibrium score regression

LoF: loss of function

LRP: Long-ranged phased

MAF: Minor allele frequency

MR: Mendelian randomization

MRI: Magnetic resonance imaging

OCD: Obesessive compulsive disorder

OR: Odds ratio

PLMS: Periodic leg movements disorder

PRS: Polygenic risk score

QC: Quality control

rg: Measure of genetic correlation

RLS: Restless leg syndrome

SNP: Single nucleotide polymorphism

TD: Tic disorder

TS: Tourette syndrome

TSQ: TS/TD screening questions

UKB: UK-biobank

WGS: Whole-genome-sequencing

List of figures

List of tables

List of original papers

This thesis is based on the following original publications, which are referred to in the text by their Roman numerals (I-V [as needed]):

I. Bertelsen, B., Stefánsson, H., Jensen, L. R., Melchior, L., Debes, N. M., Groth, C., ... **Nawaz, M.S.**, ... & Tümer, Z. (2016). Association of AADAC deletion and Gilles de la Tourette syndrome in a large European cohort. ***Biological psychiatry***, *79*(5), 383-391. [**Published**]

II. Gudmundsson, O. O., Walters, G. B., Ingason, A., Johansson, S., Zayats, T., Athanasiu, L., ... **Nawaz M.S.**, ... & Stefansson, K. (2019). Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. ***Translational psychiatry***, *9*(1), 1-9. [**Published**]

III. Yu, D., Sul, J. H., Tsetsos, F., **Nawaz, M. S.,** Huang, A. Y., Zelaya, I., ... & Tourette Association of America International Consortium for Genetics, the Gilles de la Tourette GWAS Replication Initiative, the Tourette International Collaborative Genetics Study, and the Psychiatric Genomics Consortium Tourette Syndrome Working Group. (2019). Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies. ***American Journal of Psychiatry***, *176*(3), 217-227. [**Published**]

IV. **Didriksen, M.***, **Nawaz, M. S.*,** Dowsett, J., Bell, S., Erikstrup, C., Pedersen, O. B., ... & Stefansson, K. (2020). Large genome-wide association study identifies three novel risk variants for restless legs syndrome. ***Communications biology***, *3*(1), 1-9. [**Published**]

V. **Nawaz, M. S.,** Einarsson, G., Bustamante, M., Gisladottir, R. S., Walters, G. B., Jonsdottir, G. A., ... & Stefansson, K. (2022). Thirty novel sequence variants impacting human intracranial volume. *Brain Communications*. [**Published**]

Declaration of contribution

MSN, has been involved in planning of the research work, conducted research work, performed data-analysis, drew conclusions of the study, and inlvoved in writing of the papers.

In paper I, MSN contributed to phenotype handling and identification of copy number variants, performed their association analysis to generate Table 1, Table 2, and Figure 2. MSN also participated in writing of chapters (co-author) including CNV association analysis.

In paper II, MSN contribute to search for copy number variants and provided statistical support to perform association analysis. MSN also participated in writing of manuscript (co-author) to describe association analysis.

In paper III, MSN contributed to performing the genome-wide association scan of Tourette syndrome in Iceland. MSN tested top signals from meta-analysis for replication in Iceland. MSN also perfomed polygenic risk score analysis to predict Tourette syndrome in Iceland. Altogather, MSN generated data to present Table 2 and Figure 2 of the paper. MSN, also participated in writing results chapter and methods section of the paper (co-author).

In paper IV, MSN participated in design, experiment and writing of this study as a shared 1st author with Maria D. MSN, performed meta-analysis of restless leg syndrome and subsequent analysis to generate all tables and figure of the paper. MSN analysed and wrote first version of manuscript alongwith MD and was involved in the correspondence and revision of the manuscript with journal.

In paper V, MSN was involved in conceptualizing, performing meta-analysis and subsequent analyses (genetic correlations, polygenic risk score, cis-colocalization using transcriptome and proteome data) performed in the study as a 1st author. MSN performed analyses and wrote first version of manuscript. MSN also led the internal review of the manuscript involving all authors. MSN was also in correspondence of manuscript with journal.

1 General Introduction

It took more than a decade of large sequencing efforts (Lander et al., 2001; Venter et al., 2001) and guidance from genetic maps (Augustine Kong et al., 2002) to assemble and read nature's genetic blueprint of a human being. Identification of diversity in the sequence directed the design of genotyping arrays that in turn have transformed human genetics. The large-scale generation of data by companies like deCODE genetics, and universities all over the world has improved our understanding of the contribution that genes make to the development of diseases (Emilsson et al., 2008; Gudbjartsson, Helgason, et al., 2015; Augustine Kong et al., 2002). The phenotypic variance explained by variants in the genomic sequence has been gradually increasing. As an example, for stable and easily measurable trait adult height the variance explained is ~24.6% (Yengo et al., 2018). For diseases with high heritability the explained variance can be considerable, e.g. for schizophrenia (heritability = 45.58%) the variance explained by sequence variants is ~9.0% (Calafato et al., 2018; Power et al., 2015). Thus, still only a fraction of the variance has been explained although more than 25,000 samples from schizophrenia patients have been whole genome sequenced and data from more than 200,000 patients are included in the largest meta-analysis for schizophrenia (Max Lam et al., 2019; Ripke, Walters, O'Donovan, & Consortium, 2020).

Sequence variants have been associated with educational attainment, age at first child and the number of children individuals have (J. J. Lee et al., 2018b; Mills et al., 2020). While some sequence variants are under negative selection pressure other variants are selected for. As a group, variants that associate positively with educational attainment have been under negative selection pressure in the Icelandic population during the 20th century (Augustine Kong et al., 2017) while variants that confer risk of higher BMI and variants that associate with ADHD have been increasing in frequency (unpublished data) in the Icelandic population over the same period. Changes in the environment, the impact of the industrial revolution, wars, and plagues, can shape populations altering the frequencies of sequence variants over time. Many rare variants conferring high risk of diseases are under negative selection pressure. Certain recurrent copy number variants (CNVs) confer high risk of neurodevelopmental, psychiatric disorders and negatively associate with cognitive abilities (O. O. Gudmundsson

et al., 2019; Stefansson et al., 2014; Stefansson et al., 2008). Carriers of these variants have fewer offspring, but new mutations maintain these recurrent CNVs at low but, stable frequencies worldwide (Stefansson et al., 2014).

Not only do we inherit our DNA from our parents, but our parents are also a strong environmental influence. The alleles inherited (transmitted) shape us but so do the transmitted and the non-transmitted alleles through the behaviour of our parents (Augustine. Kong et al., 2018) and several studies have furthermore suggested that childhood onset neurological disorders may derive from pre-existing intrauterine conditions or insults (Cao et al., 2006; Noonan, Haist, & Müller, 2009; Tian et al., 2006; Weng et al., 2010). Hence, the inherited alleles can directly affect risk of trait(s) through their impact on biological mechanisms while non-transmitted alleles exert their impact through parental behaviour or genetic nurturing i.e., the sequence variants impact parental traits which indirectly influence offspring's traits 'nature-nurture effect' (Augustine. Kong et al., 2018). Thus, environmental factors, including our parental genomes and their lifestyles, also contribute to the risk of developing diseases and other traits.

One of the goals in human genetic research is to identify sequence variants that are helpful as diagnostic markers. Another important goal is to find good targets, genes, and biological pathways, for drug discovery. This requires analysis of large datasets where variance in the sequence is compared to variance in phenotype. deCODE genetics, thanks to large sample sets and the long-range-phasing technology, has been successful in scanning the genome for sequence variants conferring risk of both rare diseases and common traits (Grant et al., 2006; Gudbjartsson et al., 2007; J. Gudmundsson et al., 2008; Stefansson et al., 2007; Thorgeirsson et al., 2008). Furthermore, the efforts of deCODE genetics were also successful in uncovering rare sequence variants conferring high risk of disease (T. Jonsson et al., 2012; Stacey et al., 2006; Steinberg et al., 2015; Walters et al., 2018) as well as protect against the same (T. Jonsson et al., 2012). The advancement in the full genome scanning has provided extensive information and exciting opportunities to better understand human diversity and the genetic architecture of various diseases and traits. This has widened the horizon of genetic studies both for rare (Mendelian) diseases and for common complex disorders (Hindorff et al., 2009).

Larger and larger datasets of genotyped samples from phenotyped individuals are now needed for discoveries. Large Biobanks (the UK Biobank, FINNGEN, the Estonian biobank, MVP (million veterans program from USA), 23&me, and DBDJ (the databank of Japan), to name a few) and other large samples (deCODE genetics, DBDS (the Danish blood donor study), MoBa (the mother, father, and

child cohort study) from Norway have proven most useful (Clare Bycroft et al., 2018). The largest meta-analysis to date includes millions of study subjects (J. J. Lee et al., 2018a; Nielsen et al., 2018; Kyoko Watanabe et al., 2020). While studies using relatively small samples have uncovered the first sequence variants for some traits (Stefansson et al., 2007), variants for other traits have remained elusive even though relatively large samples have been available (Arnold et al., 2018; D. Yu et al., 2019).

Through GWAS meta-analysis of five disorders and neurological traits, rare and common variants were found to associate with TS, Tics, ADHD, RLS, and human brain volume. Phenome-wide genetic correlation and polygenic risk score analyses helped uncover underlying genetic architecture of these neurological traits and identified five latent factors among these traits. Colocalization analyses implicated several genes associated with respective traits where pathway and gene-set enrichment analyses identified potential biological pathways involved in disease etiology. Bidirectional Mendelian randomization (MR) analyses of correlated disorders pinpointed causal relationship. The MR analyses using genetic variants associated with human intracranial volume (ICV) and several neurological traits revealed that ICV either has a causal effect on a neurodevelopmental disorder (ADHD) as well as on a neurodegenerative disease (Parkinson's) or confounded by closely correlated trait.

1.1 Childhood neuropsychiatric and involuntary movements disorders

Childhood neuropsychiatric disorders are complex conditions with high comorbidity with known pleiotropy (P. H. Lee, Feng, & Smoller, 2021; Z. Yang et al., 2021). The comorbidity generates cross-disorder heterogeneity that transcends diagnostic boundaries, which shapes phenotypic complexity. Such a comorbidity and heterogeneity are notable for chronic tics (Tics disorder-TD), obsessive compulsive disorder (OCD), and attention deficit / hyperactivity disorder (ADHD) which overlap three phenotypic domains: (1) involuntary urge to move, (2) impulsivity and (3) compulsive behaviour. TD also bears some phenotypic similarity with restless leg syndrome (RLS), as both are characterized by unpleasant sensation and compulsion for involuntary movement (Lesperance et al., 2004).

We applied a genetic correlation approach to better understand whether the phenotypic overlap (comorbidity) between childhood neuropsychiatric and involuntary movement disorders is also present at the genetic level. To this end, the largest available genome-wide summary data were used; from studies of five

childhood neuropsychiatric disorders (TS, TD, OCD, ADHD, and autism spectrum disorder (ASD)), two involuntary movement disorders (RLS, and Essential tremor (ET) which usually have a relatively early onset), and one late onset neurodegenerative disease with an impaired motor function component (Parkinson's disease) (see chapter 4.6: cross disorder genetic analysis). The analysis found two clusters (**Figure 1**, highlighted with black boxes) of correlations (detected through Ward's hierarchical clustering method (Murtagh & Legendre, 2014)), the first cluster showing positive genetic correlation within childhood neuropsychiatric disorders ($P < 0.05/45 = 0.0011$) and the second cluster showing positive genetic correlation between two involuntary movement disorders (RLS and ET) and a neurodegenerative disease (PD). These two clusters are joined by a nominally significant association between ADHD and RLS ($P = 0.0087$, rg = 0.18) (**Figure 1**).

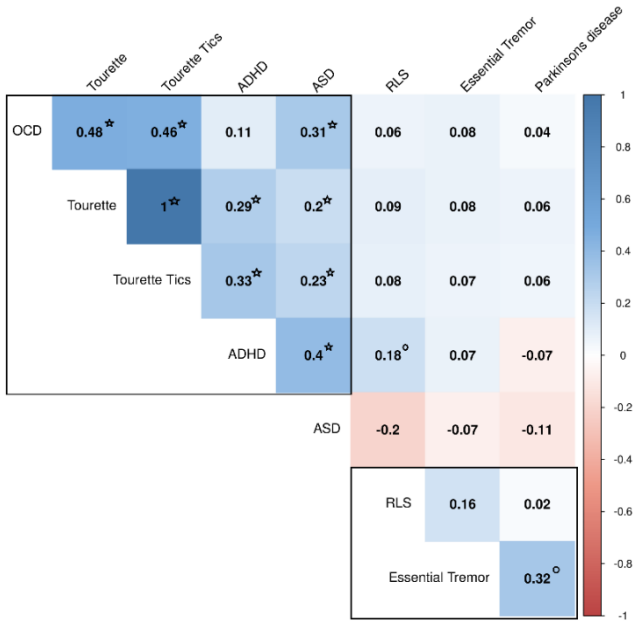


Figure 1: Genetic correlation between childhood neuropsychiatric and involuntary movement disorders. The value in each box is the genetic correlation (rg) between each pair of disorders. The Bonferroni significant associations ($P < 0.05/45 = 0.0011$) are highlighted with '☆' and the nominally significant associations as '°'. The black box bound the clusters identified through Ward's hierarchal clustering method.

Little is known about the genetic nature, etiology, heterogeneity, and the role of development in the afore mentioned conditions. Their complex and polygenic

nature requires large, genotyped study samples to uncover risk variants through genome wide association studies. The work presented here aimed at discovering sequence variants conferring risk of TD, obsessive-compulsive disorder, ADHD and RLS. These disorders are all highly comorbid with anxiety, depression, and substance use disorders (Senanayake, Krashin, & Murinova, 2020) and genetic correlation studies tell a similar story (Didriksen et al., 2020; Barbara Schormair et al., 2017). To dissect the relationship between childhood neuropsychiatric and involuntary movement disorders and their comorbidities (anxiety, depression, and life-style traits) various statistical methods were applied.

1.2 TS/TD and OCD phenotyping in Iceland

Tourette syndrome (TS) or chronic tic disorder (TD) are complex heterogenous conditions that are characterized by multiple involuntary motor and/or vocal tics. These tics are classified into 16 types (*e.g.*, facial tics, extremity tics, or vocal tics) which may differ in their manifestation. The tics may wax and wane in frequency and intensity and in some individuals they completely disappear in adulthood or through habit reversal therapy (Piacentini & Chang, 2005; Van de Griendt, Verdellen, Van Dijk, & Verbraak, 2013). TS/TD are heritable (h^2 = 0.29) (D. Yu et al., 2019), have life-time prevalence of 0.3%-1% (Brander et al., 2018; Mary M Robertson et al., 2017), and are highly comorbid with other neurodevelopmental disorders (Levy, Paschou, & Tümer, 2021; Paschou et al., 2022; Mary M Robertson et al., 2017). The heterogenous and comorbid nature of TS/TD calls for in-depth and cross-disorder analyses.

Like TS, OCD is also a complex and heterogenous disorder involving multiple obsessive and compulsive symptoms. These symptoms are characterized by recurrent, unwanted thoughts (perhaps of aggressive or sexual nature, or acts involving inappropriate behaviour in public), and repetitive behaviours. The repetitive behaviours or mental acts (such as hand washing, ordering, and checking) are performed in response to an obsession or according to rules that must be applied rigidly. They are aimed at preventing or reducing distress of a feared event or situation, a fear which at the same time is clearly unrealistic and/or excessive (Smit et al., 2020). OCD markedly impairs the quality of life by impacting personal, social, and occupational functioning and has been reported to have life-time prevalence of 2-3% (Hirschtritt, Bloch, & Mathews, 2017; Kessler et al., 2005). No unequivocal association between a sequence variant and OCD has not been reported.

In collaboration with paediatricians in Iceland, individuals diagnosed with TS/TD, and or OCD were invited to participate in the research aiming at finding risk

variants. Individuals diagnosed with TS/TD, and or OCD (ICD diagnosis) were sent invitation letters. Participants donated blood and answered screening questionnaires. Furthermore, close relatives were also invited to participate in the study. They also donated blood and answered questionnaires. As expected, there is strong comorbidity between the TS/TD sample (**Figure 1**), OCD sample (**Figure 2**) and other neurodevelopmental disorders (ADHD, and ASD).

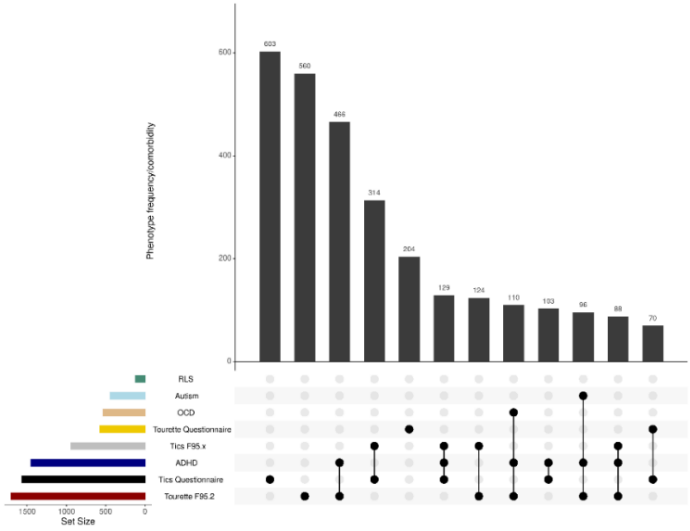


Figure 2. Upset plot showing phenotypic distribution of pure TS/TD and their known comorbidities in the studied sample. The column of dots represents the number of individuals that fulfill the criteria of the black colored dots eg. 466 individuals have both an ADHD diagnosis and a Tourette ICD10-F95.2 diagnosis

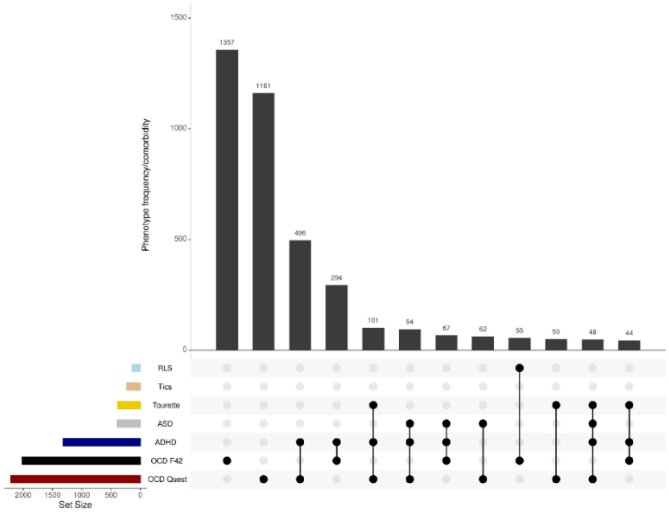


Figure 3. Upset plot showing phenotypic distribution of pure OCD and their known comorbidities in the studied sample. The column of dots represents the number of individuals that fulfill the criteria of the black colored dots, e.g., 1,357 individuals were diagnosed with OCD only (without any reported comorbidity).

1.3 RLS phenotyping in Iceland

RLS is a complex sensorimotor disorder with a prevalence as high as 10% in the general population (Maria Didriksen et al., 2017; Khachatryan et al., 2022; Pedrazzini et al., 2014; Yeh, Walters, & Tsuang, 2012). Symptoms include distressing sensations in the extremities and overwhelming urge to move the legs. These symptoms intensify when sitting or lying down. The disorder can cause reduced quality of life, poor sleep, and impaired cognitive and mental well-being (Barbara Schormair et al., 2017). Despite the high prevalence and serious health impact of the disorder, there are currently no adequate treatments for RLS as available drugs are fraught with side effects. This is in part due to limited knowledge of the pathophysiology of RLS.

In collaboration with neurologists in Iceland and Professor David Rye from Emory University, individuals diagnosed with RLS were invited to participate in a research project aiming at finding sequence variants conferring risk of RLS. Patients were initially recruited through advertisements. Participants donated blood, answered screening questionnaires for RLS and slept with leg monitors. Additionally, individuals already diagnosed with RLS (ICD 10 G25.8, available through hospital records), and their close relatives were also invited to participate in the study (Stefansson et al., 2007). In moderate-to-severely affected

subjects meeting restless leg syndrome criteria (in line with International restless leg syndrome study group rating), 5–15% do not exhibit periodic leg movements at night (PLMs) (Trotti et al., 2009). It remains unknown whether RLS in the absence of PLMs represents a separate clinical and biological entity or a limitation intrinsic to methods of ascertainment (Trotti et al., 2009). In line with Trotti et al., the PLMs screening was used to improve the diagnostic accuracy. Subjects diagnosed with RLS exhibiting PLMs ⩾ 10/h were considered confirmed RLS/PLM cases (**Figure 4**).

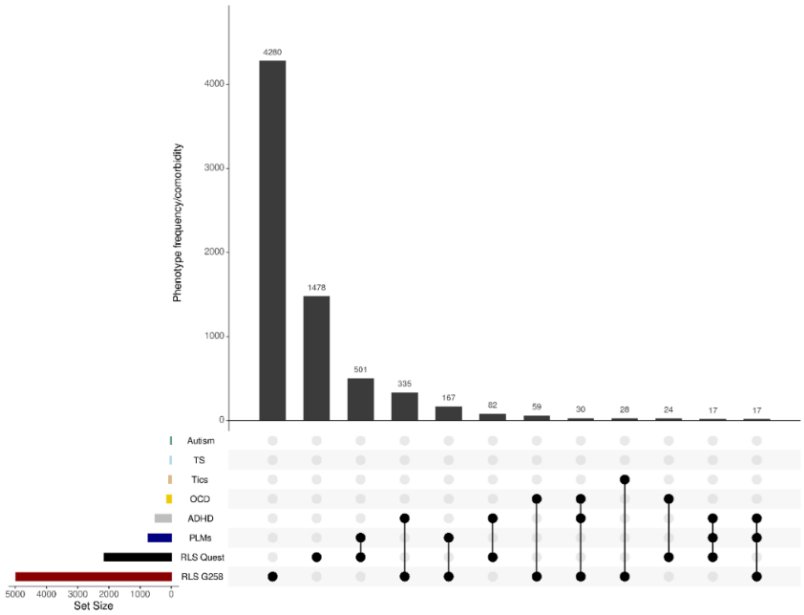


Figure 4. Upset plot showing phenotypic distribution of pure RLS/PLM and their known comorbidities in the studied sample. The column of dots represents the number of individuals that fulfill the criteria of the black colored dots, e.g., 335 individuals were diagnosed with RLS and ADHD.

1.4 DNA sequence variations

The genetic variation(s) between the DNA sequence of the individuals within the population is defined as DNA sequence variation. Spontaneous mutation (a permanent alteration to DNA sequence) or recombination events (mixing of genetic material from parents that occurrs during cross-over) are the main source of such genetic variations (Miller & Therman, 2011). The mutations can be deleterious, gain of function, or appear neutral in nature. Only those mutations

that appear in germline cells (sperms, or eggs) can be passed on to the next generation.

De novo (new) mutations are errors that occurr in the DNA replication process. These sequence variations most often are one of three types 1) single nucleotide variation, 2) insertion or deletions (indels), or 3) copy number variation. These vary in size from a single base pair change, single nucleotide polymorphism (SNP), to small size indels (1 to 100 base pairs), insertion or deletions of few base pairs, to very large structural variations (up to a few mega-bases). Copy number variations (CNVs) involve deletions, insertions, or rearrangement of chromosomal regions (Freeman et al., 2006). The CNVs introduce complex genetic variations by deleting or adding multiple genes and often impact phenotypes (Conrad et al., 2010).

The age of the father has been shown to associate positively with higher rate of *de novo* single nucleotide mutations in the offspring (H. Jónsson et al., 2017; Augustine Kong et al., 2012). The mutations vary in their impact on human traits and biological mechanisms, some with no effect, while others may provide protection against degenerative diseases (T. Jonsson et al., 2012), or cause impairments (O. O. Gudmundsson et al., 2019; Walters et al., 2018).

Sequence variants can impact gene expression and/or protein function. Mutation in coding regions of the genome affect protein function differently, synonymous variants do not change amino acid sequence and therefore unchanged protein function, while missense variants (a type of nonsynonymous variant) substitute amino acid sequence and may result in a malfunctioning protein, and finally nonsense mutations (high impact) can introduce a premature stop codon that may result in truncation or absence of a protein (through nonsense mediated decay) (Mort, Ivanov, Cooper, & Chuzhanova, 2008). Due to the high impact of coding variants, they are often in low frequency (below 1%) within a population and those with serious impact on physiology are under high negative selection pressure (frequency < 0.1%) (Reich & Lander, 2001).

1.5 Annotation of DNA sequence variants

Accurate assessment of the functional effects of sequence variants is challenging (Shameer, Tripathi, Kalari, Dudley, & Sowdhamini, 2016). The whole genome sequencing of larger samples is uncovering vast number of novel coding sequence variants that require the assessment of their functional effects. For this, researchers are using algorithms to predict the effects of amino acid changes on protein function (Adzhubei, Jordan, & Sunyaev, 2013; Sim et al., 2012). These algorithms are trained on number of the known factors (protein conservation

scores, chemical differences between amino acids) to predict the likely causal effect of coding variant. In comparison, the prediction of effect for non-coding variants (which make ~98% of the genome) is more challenging as they do not directly impact protein function. The non-coding variants (present in introns or intergenic regions) do not directly affect the protein sequence (translation) or function. They may however affect gene expression, transcript isoforms by acting through regulatory elements (Cheung & Spielman, 2009; Pagani & Baralle, 2004). One complication of gene expression analysis used to search for variants with an effect on expression (eQTLs), is that the expression and the underlying regulatory mechanism is often tissue and event specific (spatial-temporal effect). Therefore, understanding the biological effect of non-coding variants is not only complex but has also proven to be challenging (Ritchie, Dunham, Zeggini, & Flicek, 2014).

1.6 Classical genetics and candidate gene studies

Historically genetic studies were focused on Mendelian disorders by applying monogenic inheritance models. For that, linkage analysis and the candidate gene approach have widely been used (Gul et al., 2006; Nicholas et al., 2010; Santos et al., 2005). The linkage analysis (a statistical method) infers that closely located (physically) sequence variants on chromosome remain linked during the meiosis. In a family study design, this approach helps to indentify correlated segregation (linkage) of a trait and chromosomal locus (sequence variants) harboring the disease gene (Altshuler, Daly, & Lander, 2008). The linkage analysis approach has proven successful in finding highly peneterant causal genes for Mendelian diseases (Jimenez-Sanchez, Childs, & Valle, 2001). Though successful for Mendelian diseases the linkage analysis was less successful in finding loci linked to common disorders. The candidate gene- approach is a hypothesis driven search for risk variants in a biologically plausible gene, or genes. This approach ignores the genome-wide domain to search for associated variants and therefore only able to identify a fraction of genetic risk factors, but the approach has the advantage that it reduces the multiple testing burden (Hirschhorn & Daly, 2005; Tabor, Risch, & Myers, 2002). However, this approach failed in large for complex diseases because; a) sample sizes were too small, and b) marker coverage was sub-optimal by ignoring large parts of the genome. Most of the common human diseases/traits follow complex polygenic inheritance model, where multiple independent sequence variants confer small to modest risk and therefore large samples are needed to uncover statistically significant associations, and genome-wide association scans have been more useful than linkage analysis.

1.7 Genome-wide association scans

The candidate gene approach only explores a fraction of the genome while the common diseases (non-Mendelian traits) follow complex polygenic inheritance models involving multiple independent variants with biological and environmental interactions. Therefore, a more robust and hypothesis free approach is required such as genome-wide scans to find genetic variants that associate with phenotypes. The first human genome was published in 2003, and since then thousands of genome-wide association studies (GWAS) have identified thousands of sequence variants associating with diseases/traits (Bjornsdottir et al., 2019; Didriksen et al., 2020; Gisladottir et al., 2020; Grant et al., 2006; Gudbjartsson et al., 2007; J. Gudmundsson et al., 2008; B. A. Jónsson et al., 2019; Stefansson et al., 2007; Thomsen & Gloyn, 2017; Thorgeirsson et al., 2008; Visscher et al., 2017). The GWAS is a powerful statistical approach that scans millions of genetic variants to find their association with a phenotype (binary or quantitative). Only a subset of these genetic variants are directly genotyped using next generation genome-wide chip-genotyping technologies 'Illumina or Affymetrix platform' (Kennedy et al., 2003; Quail et al., 2008). To increase the number of markers available for testing the rest of the sequence variants are imputed based on chip genotypes and known correlations (linkage disequilibrium) between measured and unmeasured variants.

1.8 Phasing and imputation

Haplotypes (LD-blocks) are a combination of sequence variants at two (or more) loci that show little chance of variation during recombination in meiosis and are inherited together (Stram, 2017). Sequence variants within an LD-block show non-random correlated association of alleles with traits. GWAS studies test millions of sequence variants for association with phenotypes.

Not all the sequence variants are directly genotyped/assayed for each participant in the study. However, the statistical approaches can exploit knowledge about LD-blocks (haplotypes) to impute missing or additional sequence variants that are not directly genotyped (Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, & Peter Donnelly, 2007). The haplotype map is provided from a reference panel. The reference panel is constructed based on whole genome sequencing of a population subset (Augustine Kong et al., 2002; Augustine Kong et al., 2008). Therefore, the imputation information and accuracy of sequence variants increases with enrichment of whole genome sequencing data, *i.e.,* diverse, and high depth sequencing of participants increases haplotype information.

Prior to the imputation of additional un-typed genotypes, the haplotype phasing of existing genotype calls is the crucial and critical step which helps to perform accurate imputation (Browning & Browning, 2011). The shared haplotypes between close and distant relatives (identical by descent) can often be reliably detected (Augustine Kong et al., 2008). Icelanders are geographically isolated and a relatively homogenous population with rich genealogical records. Researchers at deCODE Genetics developed a powerful method to long-range phase genomes of Icelanders and impute sequence data efficiently to low minor allele frequencies (Augustine Kong et al., 2008). Additionally, this long-range phasing method is a powerful tool to detect recurrent mutations and to identify fine-scale recombination events.

1.9 GWAS and complex traits

The past decade has seen an explosion in the number of GWAS studies with replication data for thousands of traits. The variants discovered have helped to understand the complex genetic architecture and disease susceptibility of a wide range of behavioral, anthropometric, lifestyle, cancerous, cardio-vascular, and neurological diseases (Didriksen et al., 2020; Gisladottir et al., 2020; Grant et al., 2006; Gudbjartsson et al., 2007; J. Gudmundsson et al., 2008; J. Gudmundsson et al., 2017; Hsu et al., 2019; L. Jonsson et al., 2018; T. Jonsson et al., 2012; Lo et al., 2017; Norland et al., 2019; Thorhildur Olafsdottir et al., 2021; Stefansson et al., 2007; Styrkarsdottir et al., 2017; Styrkarsdottir et al., 2019; Thorgeirsson et al., 2008; Thorgeirsson et al., 2010; Walters et al., 2018; Zink et al., 2017). These discoveries are in addition helping to advance clinical care and personalized medicine (Tam et al., 2019).

The advancement and cost effective sequencing technologies are facilitating the inclusion of rare coding sequence variant in GWASs (Gudbjartsson, Helgason, et al., 2015). To find statistically significant associations, the multiple testing burden increases with the inclusion of rare sequence variants. For that, Sveinbjornsson et al. introduced a method to weigh the sequence variants based on their impact category i.e. a coding sequence variant is more likely to be causal than intergenic sequence variants and therefore is weighted differently than intergenic variants (Sveinbjornsson et al., 2016).

1.10 Functional annotation of associated variants

GWASs have detected thousands of genetic variants that associate with multiple traits and this number is exponentially increasing with larger meta-analysis of phenotypes involving millions of participants (Kyoko Watanabe, Taskesen, Van

Bochoven, & Posthuma, 2017). In a few cases have GWASs identified coding variants that provide insight into the causal relationship (biological mechanism) of genetic variants with traits. However, the genotype-phenotype association relationship and its causation remains poorly understood, as most GWAS signals are located in non-coding or intergenic regions, (Maurano et al., 2012). Therefore, inference of GWAS signals to biological mechanism (gene expression, gene regulation, protein function, protein-protein interaction, and biological pathways) is limited. Hence, it is of utmost importance to translate the association of genetic loci into causal variants that may help guide functional genomics experiments for drugable targets (Breen et al., 2016).

Previously, I discussed that genetic variants are correlated in the haplotype block and so their association with the traits. Therefore, the GWAS signals span a genomic region, risk locus, of multiple correlated sequence variants (Kyoko Watanabe et al., 2017). At these GWAS risk loci, some of the genes maybe relevant to disease, while others may not, but are not distinguishable just on association results alone. To disentangle and pinpoint the likely causal genes, or sequence variants, requires integration of functional information (transcriptomics, proteomics, metabolomics, and methylomics). Correlation analysis of GWAS signals with –omics data (colocalization analysis) by employing LD-block information may aid in discovering the true causal signals.

1.11 Cross trait analysis

The GWASs for the complex psychiatric disorders have so far been only moderately successful in identifying associated risk variants and causal pathways. This may partially be due to the clinical diagnosis of psychiatric disorders that have a wide range of symptoms and overlapping diagnostic boundaries. Crucially, there are no biological markers for psychiatric disorders, apart from genetic variants that are being discovered. Diagnoses are therefore made using consensus clinical criteria as inferred from the Diagnostic and Statistical Manual (DSM). Hence, the complications from comorbidity and qualitative nature of psychiatric diagnoses, compared with a disorder like hypertension or osteoporosis diagnosed by direct physical measurement, make it difficult to distinguish disease severity. Therefore, the broad symptoms and the complex cross trait phenotypic overlap may impede the identification of true biological relationship of sequence variants with specific disorders. To dissect these relationships, scientists use exploratory and confirmatory factor analysis to define the latent variables (traits). These latent traits may be used to study shared genetic architecture of disorders.

In addition, the psychiatric disorders are highly polygenic, multiple independent genetic variants confer risk. To study the polygenic nature, statistical methods are employed to construct cumulative risk scores (polygenic-risk-score) of trait A and then study its impact on other traits and vice versa. Other statistical methods used to disentangle the shared genetic architecture of disorders include genome-wide genetic correlation analysis whereby the pair-wise relationship between two (or more) disorders is studied using the resulting GWAS summary data. The pair-wise genetic correlation analysis estimates the positive or negative relationship between disorders.

1.12 Causal analysis

Besides the broadly categorized and overlapping nature of different traits, a genetic variant may also impact multiple traits through distinct pathways, pleiotropy, or through an inter-mediatory trait; genetic variant →exposure (phenotype or biological marker) →outcome phenotype i.e., phenotype i.e., Genetic variant affects either another phenotype or a biological marker which in turn impacts another trait which has a measurable/visible/interpretable effect (the outcome phenotype) (Emdin, Khera, & Kathiresan, 2017; Eriksson et al., 2017; VanderWeele, Tchetgen, Cornelis, & Kraft, 2014). The pleiotropic and causal relationship can be detected through genome-wide genetic correlation, genetic risk score analysis, or using genome wide significant genetic associations as an instrumental variable. The statistical methods (conditional association, and Mendelian randomization analysis) can be applied to study whether the genetic variant(s) associated with one trait also impact/associate with another trait. Such an analysis helps to disentangle the shared genetic architecture of two diseases and may identify intermediator traits possibly impacting both conditions through shared biological pathways. Recently, analyses of several GWAS studies showed the pleiotropic and causal relationship between different traits (Burgess, Foley, Allara, Staley, & Howson, 2020; Holmes et al., 2015; Winter-Jensen, Afzal, Jess, Nordestgaard, & Allin, 2020). Such studies have promise to identify genetic variants impacting biological pathways that are either common to both traits or distinct to one of the diseases, paving the way for personalized medicine.

2 Aims

Genetic architecture of childhood neuropsychiatric and involuntary movement disorders

Chronic tics, an involuntary movement disorder (tic disorder TD), bears some phenotypic similarity with restless leg syndrome (RLS), as both are characterized by unpleasant sensation and involuntary movement in extremities and legs (Lesperance et al., 2004). TD is highly comorbid with obsessive compulsive, and attention deficit / hyperactivity disorders (Darrow et al., 2017). Likewise, RLS is comorbid with varied psychiatric conditions that include anxiety, depression, substance (tobacco, opioid) use disorders, insomnia, and hypertension (Senanayake et al., 2020). Genetic correlation studies tell a similar story (Didriksen et al., 2020; Barbara Schormair et al., 2017). These comorbid and correlated traits have a significant impact on the quality of life.

This work aimed at uncovering sequence variants conferring risk of involuntary movement disorders (TS, Tics, ADHD, OCD and RLS), as well as search for variants affecting ICV and how ICV affects neurodevelopmental disorders on the impulsivity-compulsivity spectrum and furthermore to study cross disorder risk of common and rare variants relevant to these disorders. A further aim was to cast light on whether structural changes in the human brain cause neurological disorders or alternatively whether genetic predisposition to certain neurological or neurodevelopmental disorders impacts brain structure or development. To understand the genetic basis of brain structure and neurological disorder, brain volume GWAS meta-analysis, their genetic correlations, and bidirectional Mendelian randomization analyses were conducted.

3 Materials and methods

3.1 Phenotyping and factor analysis of TS, and TD (paper I, and III)

In Iceland, using a well characterized sample of 1,023 TS cases, a GWAS study to find sequence variants conferring risk of TS is underway. These 1,023 cases (591 with ICD 10 F95.2 and 432 recruited through questionnaire data) have been chip genotyped and long-ranged phased (LRP). LRP is a phasing method developed by Kong *et.al* (Augustine Kong et al., 2008) to correctly phase family genotype data and inform long haplotypes. Similarly, for GWAS studies of TD, ADHD, OCD and ASD well characterized phenotypes comprised of genotyped and LRP subjects (*N* = 1,048, *N* = 5,204, *N* = 575, and *N* = 540 respectively). These cases were used for polygenic risks score predictions. The cases are drawn from diagnostic registries of the main Icelandic neuropsychiatric specialty clinics to which most referrals are made.

According to the Icelandic Census 2011, the total number of persons residing in Iceland was 315,556 on 31 December of 2011. With international TS prevalence estimates between 0.3-0.9% (M. M. Robertson, Eapen, & Cavanna, 2009; Scharf et al., 2014) the index-case list represents a prevalence of 0.27%, close to the lower end of the international prevalence estimate. Studies have found that many TS cases are mild and such cases without comorbidities may not be brought to medical attention (Khalifa & von Knorring, 2003; Scharf, Miller, Mathews, & Ben-Shlomo, 2012)._ENREF_12 This study administered the brief TS/TD screening questionnaire (TSQ) (Appendix 1) based on ICD-10 and DSM-IV-TR diagnostic criteria to identify and characterize a history of TS and TD to individuals with diagnosed ASD (*N* = 266), ADHD (*N* = 280), OCD (*N* = 142), relatives of individuals with neuropsychiatric disorders (*N* = 3,286) and controls (*N* = 211). Detailed demographic statistics of all participants are presented in Appendix**, Supplementary Table 1**.

3.1.1 The TS/TD screening questionnaire (TSQ)

The brief set of TS/TD screening questions (TSQ) was designed by paediatric neurologists and clinical psychologists at the State Diagnostic and Counselling Centre (SDCC) in Reykjavik Iceland to detect current or a history of TS/TD, by

self-report (parent-report for children under 18 years of age). The TSQ was designed to reflect the diagnostic criteria used in clinical practice following both DSM-IV-TR (APA, 2000) and ICD-10 (WHO, 1992) guidelines. Appendix**, Supplementary Tables 2 & 3** show that both classification systems define these disorders similarly as mental and behavioural disorders with onset occurring in childhood or adolescence (i.e. before 18 years of age) and with comparable criteria for diagnosis (Woods & Thomsen, 2014). Relatively few substantive changes were made to diagnostic criteria for TS/TD in the updated DSM-V (APA, 2013) classification. In addition to the tic questions, surveys included a brief set of medical history and symptom-related questions regarding the main comorbidities of TS/TD; ASD, OCD, and ADHD (Cavanna, Servo, Monaco, & Robertson, 2009).

Descriptive statistics were calculated for the entire sample (*N* = 4,431) and are presented in Appendix**, Supplementary Tables 1**. Moreover, scoring rules for responses to the TSQ were established according to the diagnostic criteria (Appendix**, Supplementary Table 2**). To determine a likely diagnosis of TS, endorsement of at least two motor tics and a vocal tic with onset prior to age 18 and persisting for at least a year were required. To determine a likely diagnosis of any other TD, responses indicating a history of any persistent motor or vocal tic starting before the age of 18, not being due to another reported illness or medication, was required (See scoring algorithm and results in Appendix). For individuals with a history of tics, a motor tic count and vocal tic count was generated based on 13 motor tic symptoms and 5 vocal tic symptoms included in the TSQ.

3.1.2 Exploratory factor analysis

To perform exploratory factor analysis (EFA), correlation coefficient matrices were estimated using heterocorrelation method from ordinal responses of TSQ (using polycor, see Appendix). Therein, varimax rotation solution was used to infer factor loading and structure. Factors with eigenvalues higher than 1 were retained and characteristic consideration decided the final number of factors. To validate predicted factors, confirmatory factor analysis (CFA) was employed using psych R package (see resources). Estimation was based on weighted least-squares and minimum residual calculation. Only items having factors loading > 0.40 were retained in a factor (those with cross factor loading > 0.30 were excluded from factor analysis (FA)). Bayesian-information criterion and Tucker-Lewis index (TLI) was used as a fitness index (TLI: 0.86 - 0.97) (Appendix**, Supplementary Table 4 & 5**). CFA of tic items belonging to (1) ICD 10 F95.* and (2) questionnaire based TS/TD (screened by TSQ excluding F95.*) groups

showed that eigenvalues of three factors; body/extremity tics, facial tics, and vocal tics were greater than 1 explaining 87.12% and 92.43% cumulative variance, respectively (Appendix**, Supplementary Figure 1 & 2)**.

3.1.3 Quantitative tics and TSQ score distribution

To generate quantitative traits for tic factors, the sum of positive responses for each tic item in respective tic factor category was used. For this, sums of positive responses of 18 tics belonging to respective tic factor (as shown in Appendix**, Supplementary Figure 3**) were calculated and standardized (mean 0, SD 1) while adjusting for gender, age, and respondent type (self/parental administered). These quantitative tic traits were later used to conduct GWAS analyses, calculate their heritability and to obtain genetic correlation with TS PGRS. To understand severity of TS/TD phenotypes, the distribution of each tic factor score was assessed by comparing average standardized TSQ score for TS and TD within each recruitment group.

3.2 CNV analysis (Papers I, and II)

3.2.1 CNV calling and imputation

Detecting CNVs through chip array data is challenging, in particular calling small CNVs (Valsesia, Macé, Jacquemont, Beckmann, & Kutalik, 2013). For array-based methods a high false discovery rate is a common challenge for all available CNV prediction algorithms (Pinto et al., 2011; X. Zhang et al., 2014). Here, the long range phasing (Augustine Kong et al., 2008) of SNP array genotypes was performed to validate CNVs segregating in extended pedigrees.

The CNVs were called in a set of 150,656 genotyped and long-range phased subjects using the PennCNV algorithm (Wang et al., 2007) and the CNVs were validated using shared LRP haplotype backgrounds (Appendix**, Supplementary Figure 4**). PennCNV allows for a minimum specification of family information to increase sensitivity and accuracy of CNV calls. The inclusion of family data is, however, limited to trios and quartets with no possibilities of specifying larger sib-ships or relatives beyond first degree. In this study, the extended genealogy of the Icelandic population and the known haplotype structure was used, to validate PennCNV calls and to identify CNVs segregating in extended pedigrees on the same haplotypes. PennCNV copy number detection was performed using standard protocol (Wang et al., 2007). Allele frequencies were obtained per sample batch and adjusted for genomic waves (Diskin et al., 2008) with genotype-array specific GC-model files. Markers within the genomic super

duplicated regions described in literature (Bailey et al., 2002; Bailey, Yavor, Massa, Trask, & Eichler, 2001) were excluded, and CNVs overlapping known gaps in the assembly (UCSC Table) were also excluded prior to QC.

Chromosomes phased by LRP, and pedigree information were used to inform and verify the quality of CNV calls in the 150,656 genotyped subjects. A sliding window approach was used to identify all non-overlapping genomic segments including verified CNV breakpoints. All CNV segments that segregated in pedigrees with MAF > 0.01% were used. This gave a total of 24,053,800 CNV genotypes that map to 41,181 unique CNV bins in 134,387 subjects (Appendix**, Supplementary Table 6** & **Supplementary Table 7**). These CNVs were further imputed into 100,903 first- and second-degree relatives of 150,656 directly genotyped individuals.

3.2.2 CNVs Quality control

Sample based QC per genotyping-array was performed using the statistics from PennCNV (Wang et al., 2007). Samples were removed based on QC measures as; a) BAF-SD > mean+3SD, b) LRR-SD > mean+3SD or c) GCWF > mean+3SD. In addition, outlier samples having too many CNV calls (>mean+4SD) were removed unless > 90% of the calls were found on a single chromosome. Hence, large chromosomal CNVs were not excluded. CNV level QC was performed by excluding CNVs with < 10 SNPs/call. Adjacent calls were iteratively joined together, if the distance between calls was < 20% of the combined length (Appendix**, Supplementary Table 6**).

3.3 Meta-analysis of genome-wide association studies for restless legs syndrome (Paper IV)

3.3.1 Ethical approval of restless leg syndrome study

All participating individuals (a legal guardian in case of those below 18 years) who provided their blood and/or buccal swab sample for the genetics study of restless leg syndrome also gave written informed consents for the study.

In Iceland, the encryption of sample identifiers was performed in accordance with the regulations of the Icelandic Data Protection Authority, and the National Bioethics Committee of Iceland provided the approval of the study.

In Denmark, the participants of the Danish blood donor study (DBDS) provided written informed consent. This study was approved by The Scientific Ethical

Committee of Central Denmark (M-20090237), the Danish Data Protection agency (30-0444), and the National Ethical Committee (NVK-1700407).

In UK, all the participants of the INTERVAL dataset provided written informed consent. This study was approved by the National Research Ethics Service Committee - Cambridge East (Research Ethics Committee (REC: 11/EE/0538). The UK Biobank project is approved by the Northwest Multi-centre Research Ethics Committee, and by the Patient Information advisory Group, the National Information Governance Board for Health and Social Care, and from the Community Health Index Advisory Group. The UK Biobank also holds a Human Tissue Authority license.

In Netherlands, all the participants of the study provided written informed consent for the study of RLS. This study was approved by the Medical Ethical Committee of the Academic Medical Centre (AMC) in the Netherlands, and Sanquin's Ethical Advisory Board approved DIS-III.

In the US, all the participants of US Emory sample provided written informed consent, and an institutional review board at Emory University, Atlanta, Georgia, US, approved the study protocol (HIC ID 133-98).

3.3.2 Recruitment (restless leg syndrome)

Altogether, 480,982 participants of Caucasians ancestry (10,257 cases and 470,725 controls) from Iceland, Denmark, the UK, Netherlands, and the US were recruited in these studies. All participants of the study provided written informed consent.

3.3.3 Phenotyping of restless leg syndrome

In Iceland, a screening questionnaire was used to screen for RLS-like symptoms, both among participant recruited through a newspaper advertisement and among subjects that had participated in various studies at deCODE genetics. RLS was assessed using a questionnaire based on the International RLS Study Group diagnostic criteria (IRLSSG) (Allen et al., 2014).

For the DBDS and the INTERVAL participants the RLS status was assessed using a 10-item questionnaire with excellent diagnostic specificity (94%) and sensitivity (87.2%), 'The Cambridge-Hopkins RLS questionnaire (CH-RLSq)'. Furthermore, the definite and probable RLS cases were combined into one group and the remaining participants were included in analyses as controls.

For UK Biobank participants, the clinical diagnostic code International Classification of Diseases (ICD10), tenth revision: G25.8 was used to inform about case status of restless leg syndrome. The specific sub-code for RLS (G25.81) was not available.

For the Netherlands participants were from the Donor InSight-III (DIS-III, 2015-2016) study (Timmer et al., 2019). A self-reported questionnaire, used as part of the RISE study (Spencer et al., 2013), was used to determine RLS status. This questionnaire is based on the IRLSSG criteria and was developed in collaboration with an expert on RLS (Professor David B. Rye) (Spencer et al., 2013).

For the US Emory, a dataset from the sleep program at Emory University was included, which is a tertiary care center for RLS that is recognized as a Quality Care Center for RLS by the RLS Foundation. A clinically verified RLS affection status in this dataset was used where RLS status was assessed by one of two clinicians familiar with RLS (David B. Rye and Lynn Marie Trotti) complemented by objective measurements of periodic leg movements in sleep (PLMS) and additional secondary and supportive diagnostic features (Allen et al., 2014). For the genetics study, the analysis was limited to subject of Northern European origin in line with the participants from other populations.

For all these cohorts, the effect estimates for the 20 known RLS-associated variants (B. Schormair et al., 2017) were largely like the effect estimates observed in each of the cohorts included in this meta-analysis (**Table 2**). This indicates that the phenotypes in each cohort are comparable to previous RLS GWAS efforts. Moreover, the meta-analysis of the discovery and follow-up samples replicated 19 of the 20 previously reported variants. Of all these samples, only the dataset form UK-INTERVAL was part of previously published meta-analysis (Barbara Schormair et al., 2017).

Table 1: The questionnaire used to assess restless leg syndrome.

IRLSG RLS diagnostic criteria(R. P. Allen et al., 2014)	CH-RLSq(Allen et al., 2009)	Questionnaire used by the InSight-III cohort(Spencer et al., 2013)
1. An urge to move the legs usually but not always accompanied by or felt to be caused by uncomfortable and unpleasant sensations in the legs	Do you have, or have you had, recurrent uncomfortable feelings or sensations in your legs while you are sitting or lying down? a) Yes b) No	When you try to relax in the evening or sleep at night, how often do you have unpleasant, restless feelings in your legs that can be relieved by walking or movement? a) Never b) Rarely (2 to 4 times a month) c) Often (5 to 15 times a month) d) Very often (16 or more times a month)
2. The urge to move the legs and any accompanying unpleasant sensations begin or worsen during periods of rest or inactivity such as lying down or sitting.	Do you, or have you had, a recurrent need or urge to move your legs while you were sitting or lying down? a) Yes b) No	How often do you experience a strong urge to move your legs usually accompanied or caused by unpleasant sensations in your legs – for example restlessness, creepy-crawly, or tingly feelings? a) Never b) Rarely (2 to 4 times a month) c) Often (5 to 15 times a month) d) Very often (16 or more times a month)
3. The urge to move the legs and any accompanying unpleasant sensations are partially or totally relieved by movement, such as walking or stretching, at least as long as the activity continues.	If you get up or move around when you have these feelings do these feelings get any better while you actually keep moving? a) Yes b) No c) Don't know	Is the urge to move your legs or are the unpleasant sensations partially or totally relieved by movement such as walking or stretching? a) Yes b) No c) Don't know
4. The urge to move the legs and any accompanying unpleasant sensations during rest or inactivity only occur or are worse in the evening or night than during the day.	Are you more likely to have these feelings when you are resting (either sitting or lying down) or when you are physically active? a) Resting b) Active	Does the urge to move your legs begin, or do the unpleasant sensations begin or worsen, during periods of rest or inactivity such as when sitting or lying down? a) Yes b) No c) Don't know
5. The occurrence of the above features are not solely accounted for as symptoms primary to another medical or a behavioral condition (e.g., myalgia, venous stasis, leg edema, arthritis, leg cramps, positional discomfort, habitual foot tapping).	Which times of day are these feelings in your legs *most* likely to occur? a) Morning b) Mid-day c) Afternoon d) Evening e) Night f) About equal at all times Will simply changing leg position by itself *once* without continuing to move usually relieve these feelings? a) Usually relieves b) Does *not* usually relieve c) Don't know Are these feelings *ever* due to muscle cramps? a) Yes b) No c) Don't know If so, are they *always* due to muscle cramps? a) Yes b) No c) Don't know	At what times is the urge to move your legs or the unpleasant sensations most bothersome? a) In the morning (before noon) b) In the afternoon (before supper) c) In the evening (after supper) d) At night while sleeping e) No difference by time of day

3.3.4 Cohorts used for follow-up/replication analysis

After the discovery meta-analysis, the novel markers identified were tested for replication in two cohorts.

3.3.4.1 EU-RLS-GENE study

RLS cases in the EU-RLS-GENE study were recruited in specialized outpatient clinics for movement disorders as well as in sleep clinics in eight European countries, French Canada, and the United States. RLS diagnosis was based on a face-to-face interview by an expert neurologist, implementing the diagnostic criteria established by the IRLSSG in 2003. Ancestry-matched controls were obtained for each case sample. A total of 6,228 cases and 10,992 controls were included in the statistical analysis. Written informed consent was obtained from all participants.

3.3.4.2 US (RBC-Omics) cohort

The RBC-Omics cohort included blood donors recruited from four blood centres in the United States as a part of the Recipient Epidemiology and Donor Evaluation Study (REDS-III) (Endres-Dighe et al., 2018; Kanias et al., 2017; Yuelong Guo, 2018). RLS status was assessed using the CH-RLSq as in the DBDS and INTERVAL cohorts. Analysis in the cohort was restricted to subjects of Caucasian ancestry and included 423 cases and 7,334 controls. All subjects provided written informed consent.

3.3.5 Genotyping and Imputation analysis

3.3.5.1 Icelandic dataset

At deCODE genetics, DNA samples from 150,656 Icelanders were genotyped on one or more of 16 different Illumina SNP genotyping-arrays including 14,084 participants of the RLS study. Through whole-genome sequencing (WGS) (with mean sequencing depth of 10X, median 32X) of 8,453 Icelanders, almost 34.2 million sequence variants were identified. To increase the statistical power for the association studies, these sequence variants were imputed into the 150,656 directly genotyped Icelanders employing long-range-phasing algorithm (A. Kong et al., 2008). This generated high density SNP information haplotypes (described in detail earlier (Steinthorsdottir et al., 2016)). Subsequently, logistic regression analysis was performed for each of the imputed sequence variants accounting for cryptic relatedness and adjusting for sex and year of birth (Steinthorsdottir et al., 2016).

3.3.5.2 Danish (The Danish Blood Donor Study) dataset

DNA samples (extracted from blood) from 26,565 participants of the DBDS were genotyped using the Infinitum Global Screening Array on Illumina® genotyping platform at deCODE Genetics, Iceland. To maximize the imputation accuracy, genotyping arrays with ~660,000 common genetic markers were used, these markers span the entire genome and represent major populations. Eagle (P. R. Loh et al., 2016) was used to perform long-range-phasing employing deCODE's Northwest European (NWE) reference panel. The NWE panel was constructed through whole genome sequencing data of 15,576 individuals from Scandinavia, the Netherlands and Ireland, 8,429 Danes (1,590 of these are from DBDS). The Graphtyper (H. Jonsson et al., 2017) variant caller was used to call genotypes from whole genome sequencing data. Standard protocols for the quality control, long range phasing, and imputation of study sample were used (Steinthorsdottir et al., 2016). For the association analysis, logistic regression analysis were employed by adjusting for known confounders and cryptic relatedness as described previously(Steinthorsdottir et al., 2016).

3.3.5.3 UK (The INTERVAL Study) dataset

The UK INTERVAL samples, were genotyped using the Affymetrix UK Biobank Axiom array and the genotypes were called through the Axiom GT1 algorithm (Di Angelantonio et al., 2017). The standard quality control parameters were used for the sample and cohort level QC-analysis (i.e., excluded the samples if call rate <97%, or contamination rate >10%, or sex mismatch, or not of European

ancestry (PCA-based scores on PC1 or PC2<0). For the ancestry analysis, a set of high-quality common (MAF > 0.05), and weakly correlated (r^2<0.2 between pairs of variants) autosomal variants were used. To impute the additional autosomal sequence variants, a two step procedure was employed. To phase the genotypes, IMPUTE3 was used followed by Burrows-Wheeler transform imputation algorithm PBWT employing the UK10K and the 1000 Genomes Phase 3 reference panel (URL, https://www.internationalgenome.org/data-portal/data-collection/phase-3). The association analysis used a previously described method (Ji et al., 2017).

3.3.5.4 The UK (UK Biobank) dataset

The first set of 50,000 UK Biobank samples were genotyped using the Affymetrix UK BiLEVE Axiom array. Subsequently, the remaining 450,000 samples were genotyped using Affymetrix UK Biobank Axiom® array. These samples provided genotypes for ~850,000 sequence variants. These arrays have high content overlap, or >95% common content (C. Bycroft et al., 2018; Bycroft et al., 2017). The 1000 Genomes phase 3 (Genomes Project et al., 2015), UK10K (McCarthy et al., 2016), and HRC (Bycroft et al., 2017) reference panels were used to impute the additional genotypes in these directly genotyped subjects. The imputed genotypes were transferred to deCODE Genetics, Iceland. Therein, the sample and cohort level quality control steps were followed as previously described (Steinthorsdottir et al., 2016). Association analysis using imputed genotypes employed logistic regression with adjustment for known confounders, and cryptic relatedness (Steinthorsdottir et al., 2016).

3.3.5.5 The Netherlands (Donor InSight-III) dataset

DNA samples were genotyped for 820,967 sequence variants using the UK Biobank version 2 Axiom Array (Thermo Fisher, CA, USA) (Biobank.). After performing sample level quality control (QC) steps (i.e. call rate (≥97%), Hardy-Weinberg Equilibrium (HWE) p-value<$1x10^{-6}$, and copy number analyses (MAPD[2] value ≤0.35 and WavinessSD[3] value ≤0.1), 789,754 sequence variants were retained for the imputation and downstream analysis (I. Affymetrix, 2013, 2015; l. Affymetrix, 2017) (C. Bycroft et al., 2018). To impute genotypes for additional variants, the Sanger imputation pipeline (Eagle phasing and BWT imputation using the HRC v1.1 panel) was used (Durbin, 2014; P. R. Loh et al., 2016; McCarthy et al., 2016). Post imputation QC steps excluded rare (MAF<0.01), and poorly imputed sequence variants (imputation score R^2≤0.3). Additionally, only Caucasian samples were retained by performing ancestry check using principal components analysis (PCA) carried out in PLINK2.

3.3.5.6 US (Emory) dataset

Illumina Omni Express arrays were used for genotyping at deCODE Genetics, Reykjavik, Iceland. Sample level QC excluded markers with (<94% yield, minor allele frequency <0.1%, failed Hardy-Weinberg test ($P<1 \times 10^{-6}$), or showing significant ($P<1 \times 10^{-6}$) difference between genotype batches. To phase the QC passed genotypes, SHAPEIT (v2.790) (Delaneau, Howie, Cox, Zagury, & Marchini, 2013) was used, followed by IMPUTE2 (v2.3.2)(Howie, Donnelly, & Marchini, 2009) to impute un-genotyped variants.

For ancestry analysis, ADMIXTURE (v 1.2) (Alexander, Novembre, & Lange, 2009) and EIGENSOFT (v 6.0.1) (Price et al., 2006) were used. Based on the principal component analysis, ethnic outliers were excluded from the analysis. For the association analysis of the imputed genotypes, SNPTEST (J. Marchini, B. Howie, S. Myers, G. McVean, & P. Donnelly, 2007) was used employing the frequentist additive method while adjusting for gender, and the first twenty principal components (derived from SNP-genotypes) to correct for population structure.

3.3.6 Association analysis

The analysis included 42.9 million sequence variants (method described earlier (Saevarsdottir et al., 2020)). The genotypes of the sequence variants were estimated through LRP of haplotypes and imputation processes (Augustine Kong et al., 2008). For the quantitative traits with effective sample size of over 20,000 a linear mixed model implemented by BOLT-LMM (P.-R. Loh et al., 2015) was used to test for association between sequence variants and quantitative trait, assuming an additive genetic model. Whereas for binary trait the logistic regression analysis model was used. The quantitative traits used for analysis are standardized and follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance–covariance matrix proportional to the kinship matrix (P.-R. Loh et al., 2015). Additionally, LD score regression (B. K. Bulik-Sullivan et al., 2015) was used to account for inflation in test statistics that may arise due to cryptic relatedness and stratification. A likelihood-ratio test was used to compute all P-values, as described earlier (Benonisdottir et al., 2016). To identify a genome wide significant association, the annotation dependent significant thresholds as described in Sveinbjornsson et al were used (Sveinbjornsson et al., 2016).

3.3.7 Meta-analysis

In case of polygenic traits, individual small studies have limited power to detect sequence variants associated with phenotype of interest. For this, meta-analysis approach, combining summary statistics across independent studies, has been widely used to increase power for discovery. The meta-analysis studies are flexible to combine dozens of studies and have successfully detected significant associations using millions of individual samples (from different studies) (J. J. Lee et al., 2018a; Levey et al., 2020).

To perform meta-analysis, a few statistical approaches are used, among those the simplest is, Fisher's method to combine P-values.

$$\chi\mathbf{2} = -2 \times \sum_{k}^{i=1} \log(Pi)$$

Where k is the total number of studies, P_i is the P-value for the variant in the study i, χ2 is a chi-squared distribution with 2k degree of freedom. In case of studies having different power to detect associations, they can be weighted based on their sample size employing z-statistics.

$$Z = \sum_{k}^{i} z_i \times wi \,/\sqrt{\sum_{k}^{i} wi^2}$$

Where w_i is the square root of the sample size of the ith study, Z_i is the Z-score of the variants from the standard normal distribution defined as:

$$Zi = (sign\ of\ the\ effect\ size) \times\ \Phi^{-1}(1 - Pi/2)$$

Where Φis the standard normal cumulative distribution function (Estruch et al., 2013). In the meta-analysis, different studies may have varied power to detect associations and so should be weighted based on the effective sample size (Willer, Li, & Abecasis, 2010), which could be estimated through this formula.

$$Ne = 4/(\frac{1}{Ncases} + \frac{1}{Controls})$$

Other than the varied effective sample size, the estimated effect size varies between cohorts. Therefore, meta-analysis is performed either using fixed-effect or random-effect assumptions. Among these, the fixed-effect meta-analysis is the most common approach which assumes that the true effect for all variants is the same in all cohorts under study (fixed-effect meta-analysis). This combined effect is computed as:

$$\beta^F = \sum_{k}^{i} \beta i \times w_i / \left(\sum_{k}^{i} wi\right)$$

Where β^F is the average effect estimate which has variance:

$$var(\beta^F) = 1/\sum_k^i wi$$

Therefore, this β^F is weighted by the inverse variance of the effect estimates in each study i.e. giving greater weight to larger studies (have lower variance in the effect estimate) (Higgins & Thompson, 2002; Willer et al., 2010). However, the random effect analyses allow sizes to vary between study cohorts where weight for each study is estimated as:

$$w_i^R = 1/(1/w_i + \tau^2)$$

Where 'τ' is computed by:

$$\tau = (Q - (k-1))/(\sum_i^k w_i - (\sum_i^k w_i^2 / \sum_i^k w_i))$$

There in equation, 'Q' is Cochran's Q statistic given by:

$$Q = \sum_i^k w_i(\beta - \beta^F)^2$$

Cochran's 'Q' follows a chi-squared distribution with k-1 degree of freedom.

As compared to fixed-effect meta-analysis the random effect analyses has much lower power to detect the associations and therefore are usually not used in the discovery phase of meta-analyses. However, in case of heterogeneity and to generalize the findings it is worth testing random-effect models as well (Evangelou & Ioannidis, 2013).

The heterogeneity of effect estimates among the study cohort can be estimated using Cochran's Q-statistic from the equation above. Under the null hypothesis of no heterogeneity, it is expected to follow a chi-squared distribution with k-1 degrees of freedom. A significant deviation ($P < 0.05$) from this distribution may mean that combination of the effects under fixed effect assumption is not appropriate (Higgins & Thompson, 2002).

Population heterogeneity is also measured through I^2, which is the percentage of total variation across the studies included in the meta-analysis, that can be traced to heterogeneity rather than chance (Higgins & Thompson, 2002). I^2 is calculated from the Cochran's Q statistic in the following way (and negative values are set as 0):

$$I^2 = 100\% \times Q - (\mathrm{k} - 1)/Q$$

I^2 ranges from 0 to 100, where a value closer to 100 means more heterogeneity between the studies of the analysis. While Cochran's Q-statistic has low power for detecting heterogeneity in analyses that include a small number of studies (significance is often set at 0.1 to account for this), I^2 can be readily compared

between studies of different sizes and regardless of the effect measure (beta, odds ratio, hazard ratio etc.) (Higgins & Thompson, 2002).

When planning a meta-analysis study, there are number of guidelines for the standardization of phenotype criteria, quality controls, imputation threshold, standardization, normalization, and harmonization of effect alleles (i.e., effect estimates are standardized and normalized for same effect allele across cohorts). These QC and harmonization steps help to minimize heterogeneity that may arise due to flipped allele's or non-standarized effect estimates between studies. Additionally, critical constraint in the meta-analysis studies is to deal with the level of phenotypic heterogeneity across the cohorts which in some cases is inevitable. Phenotypic heterogeneity can arise for complex phenotypes that are difficult to define (especially questionnaire based phenotypes), for example behavioral and cognitive traits, but even when phenotypes fulfill accepted clinical criteria, heterogeneity can still arise due to population stratification or simply because genetic variants have different effects in different ancestral groups (Evangelou & Ioannidis, 2013).

For meta-analysis, GWAS summary statistics were combined using an inverse-variance weighted meta-analysis by allowing different population frequencies for alleles but assuming fixed-effects. Additionally, the heterogeneity in the effect estimates was tested using a likelihood ratio test by comparing the null hypothesis of the effect being same in both populations to the alternative hypothesis of either population having a different effect. Since the QC, imputation, and association tests, adjusting for principal components was done at cohort level and later meta-analysed together. Therefore, the joint principal component analysis of the meta-analysis set was not performed. Additionally, detailed information about the QC, imputation, and the association method for Icelandic (Styrkarsdottir et al., 2019), and UK Biobank (Astle et al., 2016) have already been described by respective cohorts.

3.3.8 Rare loss of function variants and burden analysis

The loss of function (LoF) sequence variants (frameshift, stop-codon, splice-donor, and splice acceptor) are annotated to be high impact mutations as they result in the malfunctioning of the coded protein. Largely, such mutations are found in low frequency and are hard to find and impute. An extensive whole-genome-sequencing (WGS) effort is required to uncover those sequence variants which can be prohibitively expensive. Additionally, the low frequencies of LoFs in the constrained genes, those with low probability of loss of function mutations, further limit power to detect significant associations.

To benefit from the WGS data and collectively studying the LoF mutations (simple hypothesis that LoF confer similar risk), a burden analysis approach was used (T. Olafsdottir et al., 2021). Same approach was used to report the role of *MAP1B* mutations in intellectual disability and white matter deficit (Walters et al., 2018). Furthermore, a strict threshold was defined to combine all rare loss of function mutation with minor allele frequency below 0.1% and employed rare variant burden analysis (T. Olafsdottir et al., 2021) test statistics to compute risk estimate and p-value. Top Bonferroni significant associations ($P < 0.05/N$, where 'N' is the number of genes with LoF variants below 0.1% frequency tested for association). LoF mutations in these genes were further investigated using segregation and follow-up analysis in indepdent cohort. In case of rare novel associations, genes not reported in the clincVar, were followed as constrained genes.

3.4 Genetic correlation analysis using LDSC

The genetic correlation between GWAS meta-analysis and published GWAS studies with effective sample sizes over 5,000 (N = 1,099) were performed using LDSC (B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015) ($P_{threshold} < 0.05/1{,}099 = 4.5\times10^{-5}$). Most of these GWASs were reported using UK biobank, GIANT consortium, GWAS & Sequencing Consortium of Alcohol and Nicotine, and psychiatric genomics consortium data and were accessed by downloading summary data reported by Watanabe et.al, (K. Watanabe et al., 2019) and Zhao et.al, (Zhao et al., 2019). Since most of the published GWASs used in the analysis focus on samples of Caucasian ancestry, the pre-computed LD scores from 1000 genome panel with r^2 from HapMap3 excluding HLA region (B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015) was used.

3.5 Cis-colocalization analysis of top SNPs to find eQTLs

The cis-colocalization analysis was performed to prioritize genes associated with TS, RLS variants. The analysis was performed using two approaches (1) cis-eQTL analysis from RNA expression, (2) correlation with coding variants using whole genome sequence and imputation set. For RNA sequence data, the samples from whole blood (N = 13,173), and adipose tissue (N = 686) were used from deCODE genetics (Saevarsdottir et al., 2020), and for various blood cells from exSNP database (C. H. Yu, Pal, & Moult, 2016) were queried. To claim the variants that share same causal signal (that co-localize with cis-eQTL), a strict criterion of either being the top independent cis-eQTLs from above mentioned RNA sequencing dataset or being in high LD ($r^2 > 0.8$) with our top GWAS significant associations and significant for the number of tests ($P < 0.05/N$,

where '*N*' is for number of tests) was used. Similarly, for coding variants, whether top GWAS variants are in high LD ($r^2 > 0.8$) with coding variants (missense or loss-of-function). These cis-colocalization analyses may help to implicate potential genes and so better understand their biological involvement.

3.6 Gene-based genome-wide association analysis

Complex and polygenic traits are affected by multiple sequence variants often conferring small effects. In such cases, analysis may be constrained by sample size or signals otherwise too weak to detect significant association. The aggregate analysis of multiple markers within/near the gene can help the identification of novel associations. MAGMA has implemented gene-based GWAS (GWGAS) using regression analysis of continuous properties of genes, LD structure and markers to consolidate potential signals into a single statistics (de Leeuw, Mooij, Heskes, & Posthuma, 2015). MAGMA gene-based genome wide association analysis approach was used to uncover genes associated with TS. MAGMA requires pre-computed association statistics of the sequence variants in the study.

To do so, sequence variants within 200Kb of Ensemble genes and with MAFs above 0.1% were tested for their aggregate effect with TS. The observed effects, the p-values, and the effective sample size for each variant in the study served as input for the analysis. Since TS meta-analysis was performed using Caucasian samples, the 1000genome LD structure was used to correct for population stratification, compute SNP density and other covariates used by MAGMA.

3.6.1 Pathway and gene-set enrichment analysis

To study the gene-set or pathways involved, the INRICH, DEPICT, and MAGMA based gene-set enrichment methods were used to perform tissue enrichment, and pathway analysis. For this, three approaches were used (1) using gene as the association signals employing MAGMA get-set enrichement analysis method, (2) using top associated markers (p-value threshold) employing DEPICT, (3) using tight LD-blocks of independently (r2>0.5) associated signals by employing INRICH method and performing hypergeometric computations by prioritizing candidate genes identified through cis-colocalization analysis of eQTLs and coding variants.

3.7 Causal analysis through Mendelian randomization

The Mendelian randomization (MR) analysis was performed to investigate the causal association of genetically correlated traits. For that, the GWAS significant

associations from published studies for genetically correlated traits were used. To estimate the causal effect of exposure phenotype on target phenotype, it is best to use unequivocal instrumental variables (IVs) for the analysis. Therefore, a two-sample MR approach was used to get GWAS significant associated variants (as an IV) of smoking cessation (Mengzhen Liu et al., 2019; Xu et al., 2020) (current vs former, *N* = 25), smoking initiation (Mengzhen Liu et al., 2019; Xu et al., 2020) (*N* = 387), depression (Howard et al., 2019) (*N* = 97), COPD (Sakornsakolpat et al., 2019) (*N* = 81), asthma (Olafsdottir et al., 2020) (*N* = 84), lung function (Shrine et al., 2019) (*N* = 115), BMI (Yengo et al., 2018) (*N* = 933), T2D (Mahajan et al., 2018) (*N* = 398), proxy for hypertension (systolic blood pressure, *N* = 248; diastolic blood pressure, *N* = 331; pulse pressure, *N* = 279) (Evangelou et al., 2018), stroke ischemia (Malik et al., 2018) (*N* = 31), coronary artery disease (van der Harst & Verweij, 2018) (*N* = 167), height (Yengo et al., 2018) (*N* = 3,231), and educational attainment (J. J. Lee et al., 2018b) (*N* = 1,252). These IVs were used to test for their causal effect on monocyte count. Therein, the effect estimates for all these instrumental variables (effect alleles) were looked up in the summary statistics for monocyte count association. The MR analysis was performed using R package 'MendelianRandomization' accessible from (https://cran.r-project.org/web/packages/MendelianRandomization/index.html) applying inverse-variance weighted (IVW), and MR-Egger methods. Though IVW provide robust estimates for causal effect, in cases of unbalanced pleiotropy these estimates may be biased. Therefore, the MR-Egger method was specifically employed to test whether the causal estimate by IVW is biased i.e., the intercept computed by MR-Egger is different from zero. Additionally, the diagnostic plots were generated, effect versus effect plots, funnel plots and scatter plots along with estimated regression lines. Moreover, a weighted linear regression was performed including the intercept (weighted by effect allele frequency i.e., EAF×(1-EAF)) using 'lm' method in R.

3.8 Intracranial volume meta-analysis (Paper V)

3.8.1 Phenotyping of intracranial volume

The intracranial volumes were either determined from head circumference or intracranial volume (ICV) data from the participants. These measurements were adjusted for known confounders (e.g., height, gender, age, age^2, gender×age^2), and the residuals were rank transformed, and inverse normalized to use for association studies.

3.8.2 Iceland: ICV and HC

In Iceland, the ICV data of 1,392 participants was extracted from MRI acquisitions as described earlier (Sonderby et al., 2020; Stefansson et al., 2014). These subjects participated in the various projects at deCODE genetics. The ICV data were adjusted for known confounders (Sonderby et al., 2020; Stefansson et al., 2014), the residuals were rank-transformed, and inverse normalized.

Additionally, we used manual head circumference (HC) measurements from 12,506 adults, and HC data of 1,599 children (recruited through various projects like ADHD, and ASD) were used for genetic association studies. At the recruitment centre, the HC measurements were performed as a part of a comprehensive phenotyping of the general population (the deCODE health study). For adults, the HC measurements were performed manually using a measuring tape, while the participant remained in a seated position, and each measurement was repeated three times, documenting only the largest value. Thus, the largest possible circumference was measured, from the most prominent part of the forehead above the ears to the occipital protuberance. For children, HC measurements were performed at health-care centres during a routine visit of children for developmental assessment. Hence, HC measurements were performed manually using a measuring tape, while the child rested on bed, from the most prominent part of the forehead above the ears to the occipital protuberance.

The HC measures were also adjusted for known confounders (height, gender, age, age^2, and gender×age^2) and the residuals were rank transformed, and inverse normalized. The Pearson correlation between the ICV and HC measurements is high ($N_{ICV+HC\ data}$ = 1,392, r = 0.69, $P = 6.27\times10^{-92}$) as close to reported correlation (r = 0.73, $P < 0.01$) (Hshieh et al., 2016). The residual of the inverse normalized, rank transformed and adjusted data of ICV, and HC were combined (used ICV data where both ICV and HC were available) and used as a quantitative trait to run for association analysis. All the participants (or their parents/guardian in case of minor) of the study gave written informed consent, in accordance with the declaration of Helsinki, and study was approved by the Icelandic Data Protection Authority and the National Bioethics committee (referral codes: VSN-15-241, VSN-09-098, and VSNb2015120006/03.01 with amendments, and VSN-16-093).

3.8.3 UKB: ICV

The intracranial volume (ICV) processed data of 39,283 UK Biobank participants, subset of the 500,000 UK Biobank study participants, was received for those who underwent an MRI acquisition (Alfaro-Almagro et al., 2018). After the quality control checks, outliers' removal, European ancestry filtering, and additional filtering, a final set of 37,100 subjects was retained for the final study. The ICV phenotype (Volume of estimated total intra cranial, whole brain,) was retrieved from UKB using field code '26521' as described here (Jansen et al., 2020). After the quality control criteria, the raw data were rank-transformed, inverse normalized, and adjusted for known confounders (height, gender, age, age^2, gender×age, and pc1-pc20). The residual of the inverse normalized adjusted data was used as a quantitative variable for association testing. This study was approved through UK Biobank license number 24898.

3.8.4 ENIGMA ICV + EGGC HC (head circumference):

The GWAS meta-analysis of ENIGMA ICV + EGGC HC published by Haworth et.al. (S. Haworth et al., 2019) was accessed through web-portal (link in URLs) and subsequently meta-analysed together with ICV data from Iceland and UKB.

3.8.5 Calculation of Polygenic risk score

To assess the impact conferred by the confluence of common variants, the PRS for each of the 500,000 UK Biobank subjects, and 150,656 participants from deCODE study subjects in Iceland was derived. For each population, the PRSs were constructed using the GWAS summary statistics of a trait excluding the data from same population (to avoid inflation due to population biasness). Briefly, the 630,000 informative SNPs (which tag almost all LD panels in the genome) were used by weighing their effects (from GWAS summary) through LDpred, as described previously (Vilhjálmsson et al., 2015). The following formula was used to calculate the PRS score from the weighted data.

$$PRS_j = \sum_{i \in s} w_i \times G_j$$

Where '*s*' is the set of genetic variants retained under p-value threshold, 'w_i' is the weight given to the ith variant, given by the log odds ratio or regression coefficient in a regression (the GWAS study from which PRS score is to be calculated), 'G_{ij}' is the expected count of effect allele in the individual 'j'.

Subsequently, for sanity check and to find the best weight, the constructed PRS score was used to assess their impact on respective phenotype i.e., the weight which explained highest phenotypic variance.

This weight was further used to perform phenome wide PRS association analysis to study impact of PRS on other traits. To this end, for binary traits the logistic regression method was employed using population controls that fall in the same year of birth bin and by adjusting for gender, poly(year of birth,4), and the first 40 (PC1-PC40) principal components derived from the SNP genotypes. For quantitative traits, the linear regression method was employed by adjusting for above mentioned covariates. The estimated p-values for each trait were further corrected for inflation using inflation factor computed by LDSC (B. K. Bulik-Sullivan et al., 2015). To report a significant association, a strict Bonferroni significant p-value threshold was used ($P < 0.05/N$), where 'N' is the number of traits tested for association.

4 Results

4.1 Genetics of Tourette syndrome

Tourette syndrome (TS) is a complex heterogeneous disorder characterized by motor and vocal tics. The tics may wax and wane in frequency and intensity. Also, in some individuals the tics may completely disappear in adulthood. TS is highly comorbid with three other neuropsychiatric disorders, obsessive-compulsive disorder (OCD), attention deficit hyperactive disorder (ADHD), and autism spectrum disorder (ASD). The heterogeneous and comorbid nature of TS requires in-depth and cross-disorder analysis to better understand its biological nature. The GWAS search for common and rare variants including CNVs associating with TS was performed. As a part of TS-EUROTRAIN network, the largest European GWAS meta-analysis for TS was planned to better understand the aetiology and pathophysiology of the disorder (Forde et al., 2016).

In collaboration with clinicians at Landspitali University hospital in Iceland, the 2,023 individuals with TS ($N_{F95.2}$ = 1,632), and 1,322 with TD ($N_{F95.0,\ F95.1,\ F95.8,\ F95.9}$ = 651) were identified. The recruitment centre (Þjónustumiðstöð rannsóknarverkefna), a clinic administered under conditions issued by the Data Protection Authority of Iceland, contacted the affected individuals or their legal guardians if the subjects were younger than 18 years of age and offered them participation in the study. All participants who donated samples gave written informed consent and the National Bioethics Committee of Iceland approved the study.

A subset of clinically diagnosed TS/TD cases, their relatives, and an additional set of subjects (who also participated in other studies conducted by deCODE genetics) participated by answering questionnaires, providing data for diagnosis and detailed tics phenotyping (N = 14,633). The questionnaire data were used to perform factor analysis to better understand the phenotypic heterogeneity and to generate quantitative tics phenotypes. Based on tics-questionnaire data, the analysis detected three latent tic components that are clustered as (1) vocal tics (2) facial motor tics, and (3) extremity and abdominal region motor tics (**Figure 5**). Loading of these clusters was consistent in clinically diagnosed and questionnaire-based TS and TD cases (**Figure 5**).

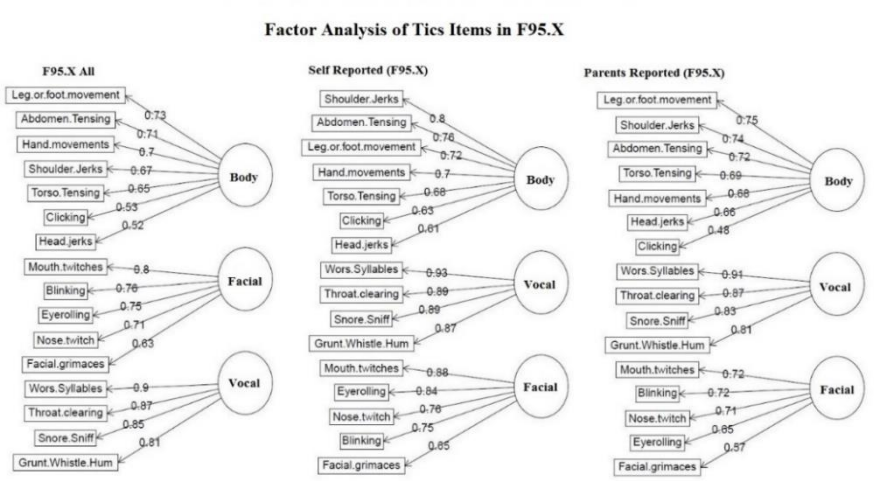


Figure 5: Factor analysis of tics questionnaire data. **Factor analysis of tic types in clinical (ICD 10 F95x) cases using heterochor correlation analysis**. To understand tics heterogeneity, the recruited subjects diagnosed with Tourette and/or tics (*N* = 930) were assessed for the prevalence of either of the 16 tic types. Factor analysis of those response resulted in three tic factors (a) body/extremity tics (b) facial tics and (c) vocal tics.

4.2 Copy number variations (CNVs) analysis of TS, and ADHD (Paper I, II and unpublished data)

We carried out three CNV analyses. First, candidate CNVs reported to associate with TS (in scientific publications) were studied in the European sample described previously (1,181 TS cases, and 118,730 controls). A second set of 19 neuropsychiatric CNVs conferring risk of Autism and Schizophrenia were tested for their association with ADHD and TS in Icelandic and Norwegian samples. Third, genome wide CNV analysis was carried out for TS.

4.2.1 Candidate CNVs study of TS (Paper I)

For TS some genome wide CNV studies have suggested involvement of a few CNVs in the pathogenesis of TS (Fernandez et al., 2012; McGrath et al., 2014; Nag et al., 2013; Sundaram, Huq, Wilson, & Chugani, 2010). In a small study, Sundaram et al. (Sundaram et al., 2010) reported four recurrent CNVs deleting exon(s) of *NRXN1, CTNNA3, FSCB*, and *AADAC*. Partial deletions in *NRXN1* and *CTNNA3* have previously been associated with ASD (Marshall et al., 2008; Wang et al., 2009) and/or schizophrenia (Kirov et al., 2008). Sundaram et al. found suggestive association of *AADAC* deletion with TS (Sundaram et al., 2010).

To study these candidate genes, Danish collaborators initially screened 243 TS cases and 1,887 matched controls from Denmark. While the association was not significant in Denmark (P = 0.13, OR =1.43) it prompted further studies. The association signal of *AADAC* deletion was followed in five additional European populations including Iceland, Netherland, Hungary, Germany, and Italy. The Mantel-Haenszel meta-analysis of 1,181 cases and 118,730 population controls confirmed the association with the *AADAC* deletion ($P = 4.4\times10^{-4}$, OR = 1.90) and has been described (Paper I) (Bertelsen et al., 2016).

In addition, a Norwegian sample has now been genotyped, and analysis of the data further supports the association between *AADAC* and TS/TD (ICD10 F95, **Figure 6**) (unpublished data, collaboration with Prof. Ole Andreassen). Moreover, using additional 669 TS cases from Iceland, the combined meta-analysis of European samples is consistent in the larger sample ($P = 1.7\times10^{-5}$, OR = 1.58, **Figure 6**).

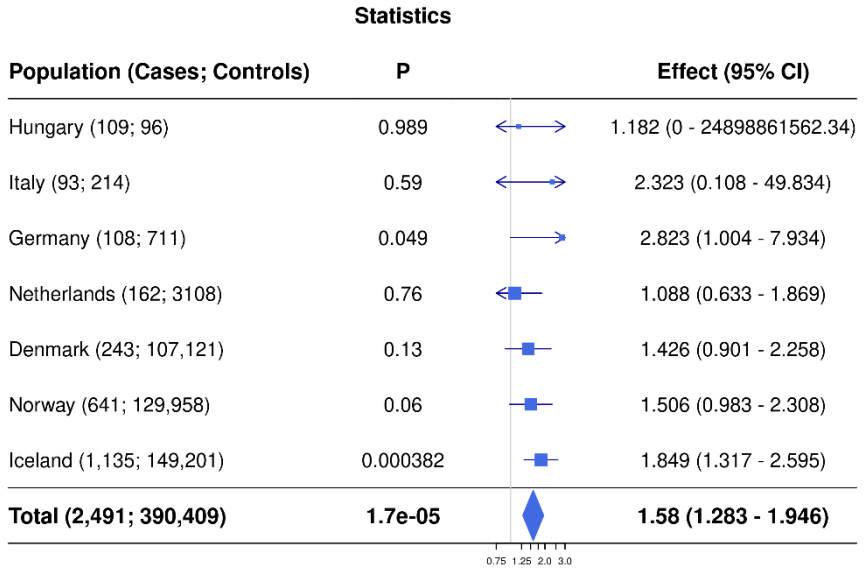


Figure 6. Meta-analysis of *AADAC* CNV deletion including Norwegian samples and additional samples from Iceland. In the follow up study, the *AADAC* deletion was tested in the larger Icelandic and additional Norwegian samples. Each blue line indicates effect size in respective population, and blue diamond represents combined effect 'updated meta-analysis' of *AADAC* deletion.

A panel of total RNA from 19 different regions of human adult brain was used to study the expression of *AADAC* in the central nervous system. Expression of *AADAC* was found in all the 19 regions.

4.2.2 Neuropsychiatric CNV analysis in ADHD (Paper II)

Rare recurrent CNVs have been associated with neurological disorders such as schizophrenia, ASD, developmental, and neuropsychiatric disorders (Ingason et al., 2011; Kirov et al., 2014; Malhotra & Sebat, 2012; Morrow, 2010; Stefansson et al., 2008). A group of 19 large and rare CNVs (0.0027–0.25% carrier frequency in the population) referred to as 'neuropsychiatric-CNVs' also affect cognition and negatively impact educational attainment (Stefansson et al., 2014). Two of these 19 neuropsychiatric CNVs were reported to associate with ADHD (Schneider et al., 2014; Williams et al., 2010). Paper 2 represents the largest study to test a set of neuropsychiatric CNVs with ADHD. Furthermore, the same variants were tested for association with TS. The analysis found that neuropsychiatric CNVs are in higher frequency in ADHD and TS than in population controls.

Through meta-analysis of Icelandic and Norwegian data, the 19 neuropsychiatric-CNVs as a group confer risk of ADHD ($P = 1.6\times10^{-21}$, OR = 2.43) (O. O. Gudmundsson et al., 2019). These CNVs are found in low frequency in the population but as a group they are found in 1.51% frequency. In this sample, 14 of the 19 CNVs were tested for association (have sufficient power to detect association with CNV frequency > 0.018%) between ADHD and population controls (**Figure 7**). The analysis identified six false discovery rate adjusted significant CNV associations ($P < 0.05/14 = 0.0036$) including the 22q11.21 deletion ($P = 1.8\times10^{-6}$, OR = 10.73), the 16p11.2 proximal duplication ($P = 9.1\times10^{-5}$, OR = 4.34), the 15q13.3 (BP4 & BP4.5-BP5) deletion ($P = 1.0\times10^{-4}$, OR = 5.97), the 1q21.1 distal duplication ($P = 0.0020$, OR = 3.44), the 2p16.3 *NRXN1* deletion ($P = 0.0026$, OR = 4.68), and the 16p13.11 duplication ($P = 0.0035$, OR = 2.12) (**Figure 7**) (O. O. Gudmundsson et al., 2019).

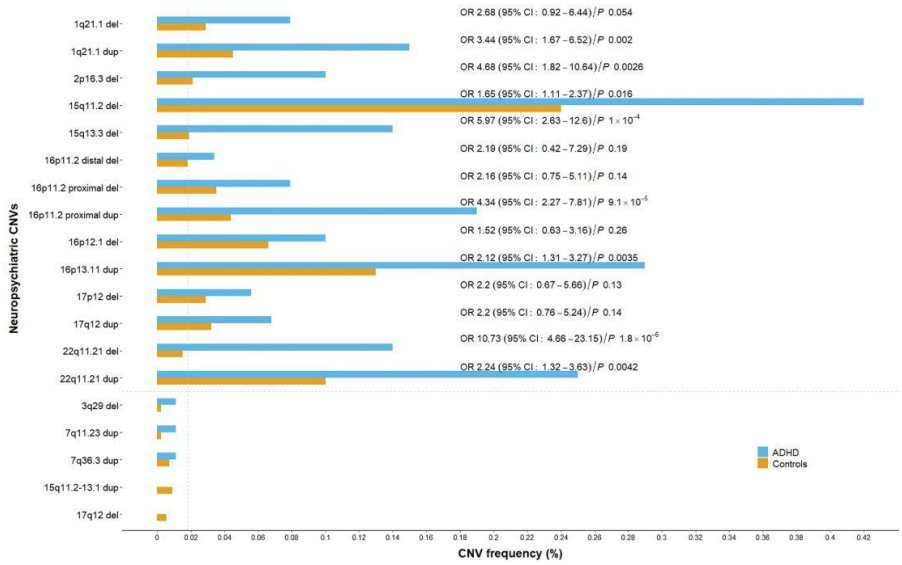


Figure 7. Summary of 19 neuropsychiatric CNV associations with ADHD in Icelandic and Norwegian samples. The carrier CNV frequency for ADHD cases and controls was calculated from a combined Icelandic and Norwegian dataset. The Cochran-Mantel-Haenszel χ^2 test was used for count data to estimate OR and *P* value from a combined set of Icelandic and Norwegian genotypes. The effects were adjusted for cryptic relatedness and population structure using the intercept from LD score regression (B. K. Bulik-Sullivan et al., 2015).

4.2.3 Neuropsychiatric CNV analysis of TS (unpublished data)

As TS and ADHD are highly comorbid disorders with an early age of onset (Hirschtritt et al., 2015). It is known that neuropsychiatric CNVs have pleiotropic effect on ADHD, autism, and schizophrenia (O. O. Gudmundsson et al., 2019; Malhotra & Sebat, 2012; Rees et al., 2014; Rees, O'Donovan, & Owen, 2015) but their effect on TS/TD is not explored yet. We used a combined sample of Icelandic and Norwegian TS/TD cases (N = 2,684) and controls (N = 279,143) and tested 13 of the 19 neuropsychiatric CNVs. At a $P_{threshold}$ < 0.0039 (0.05/13), we found that the 17q12 duplication confers high risk of TS/TD (P = 4.4×10^{-5}, OR = 10.22, **Table 2**). The 17q12 duplication was not reported to associate with ADHD (P = 0.14, OR = 2.20) (O. O. Gudmundsson et al., 2019). To replicate these findings, this CNV was tested in an independent Danish sample of 246 clinically diagnosed TD (F95.x cases) and 94,363 population controls. While the risk (Odds ratio) is in keeping with the discovery observation, the replication is not significant (P = 0.08, OR = 11.79). The combined meta-analysis of 17q12 duplication using Icelandic, Norwegian, and Danish data

further strengthen these findings (***P* = 8.7×10^{-6}, OR = 10.43**). Additionally, the psychiatric CNVs as a group also confer risk of TS/TD (*P* = 0.0011, OR = 1.90). The association of 17q12 duplication and neuropsychiatric CNVs with TS/TD, highlight its shared component with other neuropsychiatric disorders. The 17q12 duplication has also been reported to negatively associate with performance on tests for cognitive function (Stefansson et al., 2014). A more detailed analysis is required to further understand how 17q12 duplication affects TS/TD.

Table 2. Effect estimates for association testing of neuropsychiatric CNVs with TS/TD. OR is estimated odds ratio, *P* is the p-value for the association test, F1 is the frequency (in percentage) of CNV in cases whereas F2 is frequency (in percentage) of CNV in controls.

Psych_CNV	Iceland			Norway			Combined		
	Tourette + Tics (Cases = 2,043; Controls = 149,201)			Tourette + Tics (Cases = 641; Controls = 129,942)			Tourette + Tics (Cases = 2,684; Controls = 279,143)		
	F1/F2	*P*	OR	F1/F2	*P*	OR	F1/F2	*P*	OR
1q21_1_distal_dup	0.049/0.048	0.989	1.014	.	.	.	0.049/0.048	0.989	1.014
2p16_3_NRXN1_refseq_del	0.049/0.019	0.326	2.609	.	.	.	0.049/0.019	0.326	2.609
15q11_2_del	0.196/0.241	0.998	0.811	0.156/0.068	0.355	2.306	0.186/0.161	0.36	2.31
15q13_3_BP4_BP45_BP5_del	0.098/0.018	0.058	5.415	0.156/0.006	0.043	25.388	0.112/0.013	0.0083	7.81
16p11_2_distal_del	0.049/0.017	0.307	2.812	.	.	.	0.049/0.017	0.307	2.812
16p11_2_proximal_del	0.049/0.033	0.493	1.491	.	.	.	0.049/0.033	0.493	1.491
16p11_2_proximal_dup	0.098/0.042	0.219	2.321	.	.	.	0.098/0.042	0.219	2.321
16p12_1_del	0.147/0.064	0.147	2.308	0.156/0.021	0.129	7.518	0.149/0.044	0.053	2.78
16p13_11_dup	0.098/0.121	0.998	0.807	.	.	.	0.098/0.121	0.998	0.807
17p12_del	0.049/0.029	0.45	1.699	.	.	.	0.049/0.029	0.45	1.699
17q12_dup	**0.147/0.03**	**0.027**	**4.874**	**0.468/0.013**	**1.26e-4**	**35.968**	**0.224/0.022**	**4.4e-05**	**10.22**
22q11_21_17MB_20MB_multiple_dup	0.049/0.019	0.335	2.519	.	.	.	0.049/0.019	0.335	2.519
22q11_21_17MB_20MB_multiple_del	0.049/0.019	0.335	2.519	.	.	.	0.049/0.019	0.335	2.519
Combined (Neuropsychiatric_CNVs)	**1.224/0.774**	**0.0458**	**1.523**	**0.936/0.108**	**9.26e-5**	**8.761**	**1.155/0.464**	**0.0011**	**1.90**

4.2.4 SNP GWAS meta-analysis for Tourette (Paper III)

Paper III reports the GWAS meta-analysis of 8.3 million sequence variants tested for association with TS in 4,819 cases and 9,488 matched controls of European ancestry. In this study the Icelandic TS sample was used as a follow up sample to test for top sequence variants while PRS score for TS was constructed for 150,250 Icelanders to study PRS distribution in population.

The discovery meta-analysis yielded one GWAS SNP (*P* = 2.1×10^{-8}, OR = 1.16, **Figure 8**). However, the association signal did not replicate in the Icelandic sample. The discovery signal is an intronic variant at 13q12.2 in *FLT3*. Thirty-nine independent signals (MAF > 1.00%, and *P* < 1.0×10^{-5}) from discovery meta-analysis were tested for replication in the deCODE sample. None of the discovery signals were confirmed in the Icelandic sample (*P* > 0.05/39 = 0.0013). Summary statistics for top 10 signals in the primary analysis are shown in **Table 3**. TS PRS analysis highlighted TS polygenicity where TS PRS scores were elevated in TS, and TD compared to population and screened controls (**Figure 9**).

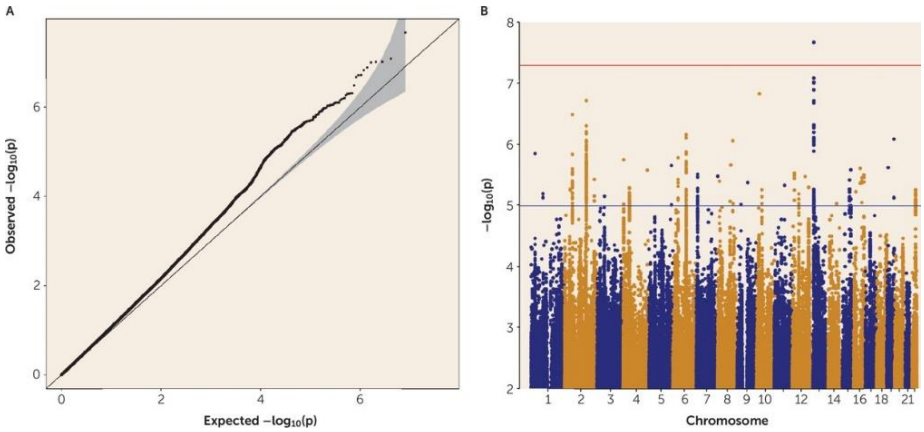


Figure 8: Results of the primary Tourette's syndrome genome-wide association study meta-analysis of 4,819 cases and 9,488 controls. Panel A is a quantile-quantile plot of observed versus expected -log10 (p) values from the primary genome-wide association study (GWAS) meta-analysis. The 95% confidence interval of expected values is indicated in grey. The genomic control λ value is 1.072, and the λ1000 value is 1.011 for single-nucleotide polymorphisms (SNPs) with minor allele frequency >0.01, INFO score (measurement of imputation quality) >0.6, and certainty >0.9. Panel B is a Manhattan plot of all final genotyped and imputed SNPs in the primary Tourette's syndrome GWAS meta-analysis. The upper horizontal line indicates the genome-wide significance threshold of 5×10^{-8}, and the lower horizontal line indicates the suggestive threshold of 1.0×10^{-5}.

Table 3: Top 10 linkage disequilibrium–independent loci in the primary Tourette's syndrome GWAS meta-analysis. Chr is for chromosome, Pos_hg19 is position in hg19, rsID is the dbSNP ID of the variants, EA is effect allele, OA is other allele, EAF% is effect allele frequency in percentage, Gene is the closest genes within 500Kb, P is p-value for association, OR is estimated odds ratio for the effect allele. [a] P and OR is presented for the primary Tourette's syndrome GWAS meta-analysis of 4,819 cases and 9,488 controls, [b] for the targeted replication in the independent deCODE sample (706 Tourette's cases and 6,068 controls), and [c] for the meta-analysis of these two data sets.

Top 10 linkage disequilibrium–independent loci in the primary TS							Primary[a]		Follow up[b]		Combined[c]	
Chr	**Pos_hg19**	**rsID**	**EA**	**OA**	**Gene**	**EAF%**	**OR**	***P***	**OR**	***P***	**OR**	***P***
chr1	29576784	rs6670211	A	C	*EPB41*	42.06	0.88	1.4e–06	0.94	0.45	0.89	1.5e–06
chr2	161544891	rs13407215	T	C	*AHCTF1P1*	0.01	2.21	1.9e–07	0.02	0.85	2.21	1.9e–07
chr2	58955953	rs2708146	G	A	*LINC01122*	48.05	0.88	3.2e–07	0.98	0.75	0.89	8.0e–07
chr4	2460571	rs73205493	T	C	*LOC402160*	35.09	1.16	1.8e–06	1.08	0.34	1.15	1.6e–06
chr6	36623338	rs72853320	A	G	*CDKN1A*	12.13	1.2	1.7e–06	0.88	0.28	1.17	2.2e–05
chr6	98550289	rs1906252	A	C	*MIR2113*	50.13	0.88	7.0e–07	0.9	0.17	0.88	2.8e–07
chr8	113581898	rs117648881	A	G	*CSMD3*	1.13	0.59	8.8e–07	0.72	0.32	0.6	6.2e–07
chr10	23705451	rs191044310	A	T	*OTUD1*	0.24	0.54	1.5e–07	2.27	0.25	0.56	5.9e–07
chr13	**28612886**	**rs2504235**	**A**	**G**	***FLT3***	**32.02**	**1.16**	**2.1e–08**	0.94	0.5	1.14	2.4e–07
chr19	52318380	rs12459560	T	G	*FPR1*	16.17	1.19	8.2e–07	1.08	0.45	1.18	9.1e–07

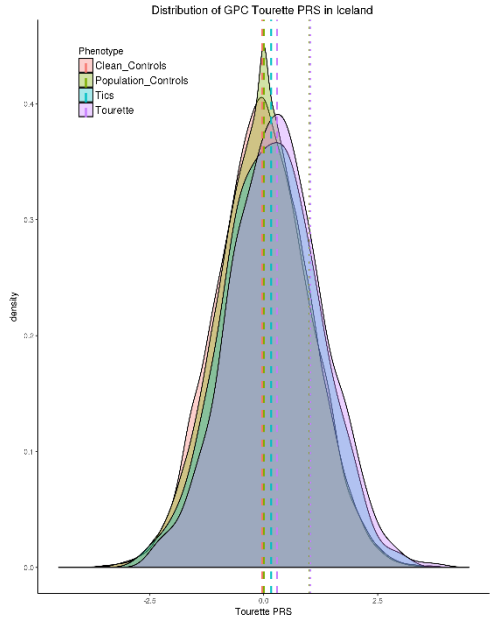


Figure 9: TS Polygenic risk score density plot in population-based sample from Iceland. The plot shows PRS in TS cases (*N* = 706), TD cases (*N* = 466), unscreened population control subjects (*N* = 127,164), and tic-negative control subjects (*N* = 6,068).

4.3 GWAS analysis of TS, and TD including rare variants (unpublished data)

The largest TS GWAS, to date, includes 4,819 cases and 9,488 controls (D. Yu et al., 2019). Due to limitations in imputation of rare variants, current GWAS studies cover only common variants (MAF above 1%). To date no SNPs or Indels have been unequivocally associated with TS or TD (D. Yu et al., 2019).

In Iceland, approximately half of the adult population has been genotyped using Illumina SNP arrays. Furthermore, more than 15% of the population has been whole genome sequenced. This allows for long-range phasing of the Icelandic chromosomes and for imputing rare as well as common variants into the phased chromosomes (Gudbjartsson, Sulem, et al., 2015; Augustine Kong et al., 2008). Hence, in the Icelandic sample, variants with a MAF as low as 0.01% can be tested for association with diseases and other traits.

Under the additive model, 42.9 million sequence variants were tested for association with TS and TD in the Icelandic sample. The tested variants were identified through whole-genome sequencing of 49,708 Icelandic individuals. The variants were subsequently imputed into long range phased chromosomes of

166,281 chip genotyped Icelanders, as well as 150,998 of their close relatives (Gudbjartsson, Helgason, et al., 2015). Weighted Bonferroni thresholds for sequence variant annotation classes were applied to determine significance (Sveinbjornsson et al., 2016).

The GWAS for broad TS (cases = 2,023; controls = 317,445) did not yield any significant associations (**Figure 10A)**. Furthermore, the top 10 associations ($P < 5\times10^{-8}$) were not confirmed in a Norwegian follow up sample (**Table 4**).

The GWAS for broad TD (1,322 cases; 326,339 controls) identified one genome-wide significant variant (rs564796941-C, $P = 4.52\times10^{-13}$, OR = 5.05, EAF = 0.33%, **Figure 10B and Table 4**). rs564796941-C is a frameshift mutation (p.Lys271ThrfsTer37) predicted to be a stop-gain, loss of function variant in the epidermal growth factor-like 7 (*EGFL7*). This mutation is present in the last exon (exon 13) of *EGFL7* and so it is not clear what impact its presence has on the gene product; stop-gain mutations in the final exon are not always detrimental where early truncation does not have a negative impact and the function of the protein may be preserved. We searched for other loss-of-function sequence variants in *EGFL7* and found a rare splice acceptor (rs755785262-G) and a splice donor (rs746089480-A) that were tested for association with TD (**Figure 11**). These variants do not individually associate with TD (**Table 4,** rs755785262-G, $P = 0.113$, OR = 2.84, EAF = 0.08% and rs746089480-A, $P = 0.332$, OR = 1.39, EAF = 0.41%, respectively). A burden test (T. Olafsdottir et al., 2021) of these three potential loss-of-function variants (1,177 cases; 165,104 controls) shows significant association with TD ($P = 7.76\times10^{-10}$, OR = 5.14).

The frameshift variant, rs564796941-C, is rare in the Norwegian sample (ICD-10 F95 cases = 399, controls = 67,533). While the OR estimate is in keeping (OR = 4.15) the association is not significant ($P = 0.29$) and a larger sample is needed to unequivocally confirm or reject this association. The splice acceptor (rs755785262-G) and splice donor (rs746089480-A) variants were not found in the Norwegian sample.

The predicted loss-of-function sequence variants (frameshift variant, introduction of stop codon, and splicing variants) may impact gene expression differently. The *EGFL7* variant associating with TD, rs564796941-C (p.Lys271ThrfsTer37), does not associate with *EGFL7* expression (**Figure 11B**) although it adds additional 37 amino acids. rs746089480-A (not in LD with rs564796941-C), another variant predicted to be loss-of-function in *EGFL7*, has a similar frequency to rs564796941-C, but does not associate with TD. This variant is on a haplotype background that is associated with a high expression of *EGFL7* and is in high LD

(r^2 = 0.83) with the second leading eQTL of *EGFL7* (**Figure 11B**). The rs746089480-A causes skipping of exon 11 of *EGFL7* ($P = 2.02\times10^{-95}$, βexp = 2.54). The third variant, rs755785262-G, predicted to be a loss-of-function in *EGFL7*, is found in lower frequency than the other two and does not associate with TD or *EGFL7* expression (**Figure 11B**). Hence, based on RNA expression analysis alone, it can't be determined whether these *EGFL7* sequence variants have loss or gain of function state.

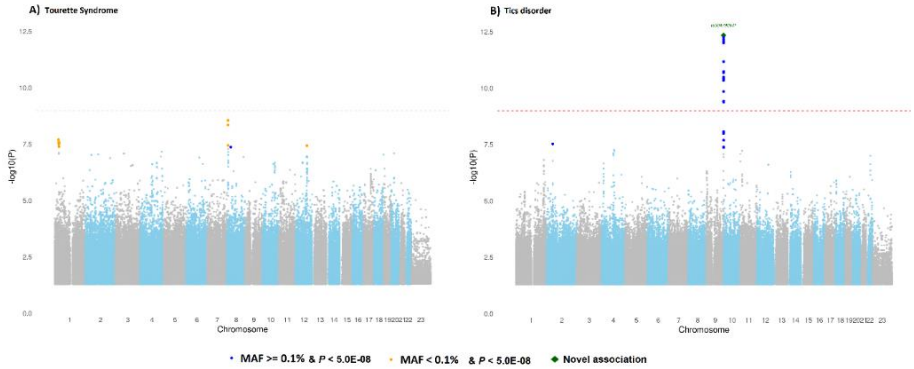


Figure 10: SNP GWAS Manhattan plot for (A) Tourette syndrome, (B) Tics disorder. The red dotted line represents weighted significant threshold based on variant annotation (Sveinbjornsson et al., 2016) ($P < 0.05/42.9\times10^{-7} = 1.02\times10^{-9}$). Variants with ($P < 5.0\times10^{-8}$) are labeled orange if EAF < 0.1% and blue if EAF >= 0.1%. Variants with a $P < 0.05$ were used to generate the Manhattan plots. In the plots Chromosome 23 refers to chromosome X.

Table 4: Summary data for top sequence variants associated with TS, and TD. Chr is for chromosome, Pos is position in hg38, rsID is the dbSNP ID of the variants, *P* is the p-value for association, OR is the estimated odds ratio for the effect allele. Phet is p-value for heterogeneity test, I2 is percentage of variation across cohorts that is due to heterogeneity rather than by chance. EA is the effect allele, OA is for other allele, and EAF is the effect allele frequency in percentage. Gene is the closest gene within 500Kb, category is the category of the effect allele type. [a] Is for discovery sample from Iceland. [b] Is for follow up sample from Norway. [c] combined meta-analysis of discovery + follow up sample. Cases is for number of cases used, and controls is for number of controls used in the analysis.

Top sequence variants associated with TS, TD, and TS+TD using Icelandic discovery sample							Discovery[a]			Follow up analysis[b]		Combined analysis[c]						
Chr	**Pos_hg38**	**rsID**	**EA**	**OA**	**Gene**	**category**	**EAF%**	***P***	**OR**	***P***	**OR**	***P***	**OR**	**95%CI**	**Phet**	**I2**	**Cases**	**Controls**
Tourette Syndrome (TS)																		
chr4	140524735	rs548278560	G	A	*ELMOD2*	intron variant	0.126	7.26e-06	4.819	0.6347	0.018	8.1e-06	4.77	(2.40,9.48)	0.51	0	2,283	384,969
chr8	29832791	rs142477257	C	G	*LOC101929470*	regulatory region variant	1.836	2.89e-07	0.293	0.4212	0.634	4.71e-07	0.33	(0.21,0.51)	0.21	36.5	2,283	384,969
chr11	104954870	rs56226603	T	C	*CASP4*	missense variant	0.002	5.65e-06	87.331	0.6085	0.018	9.0e-06	76.55	(11.28,519.57)	0.28	13.3	2,283	384,969
chr12	87088607	rs766005749	T	C	*LOC105369878*	intergenic variant	0.012	8.03e-06	21.025	0.31558	0.017	2.3e-05	17.29	(4.63,64.64)	0.084	66.6	2,283	384,969
chr12	90673999	rs777704260	C	T	*LOC102724834*	intergenic variant	0.033	5.74e-07	11.870	0.51608	0.018	8.1e-07	11.39	(4.33,29.95)	0.30	8.6	2,283	384,969
chr12	91297554	rs146533880	C	T	*LOC105369898*	intergenic variant	0.015	3.07e-07	22.396	0.23703	2.875	9.7e-07	11.70	(4.37,31.32)	0.057	72.3	2,283	384,969
chr13	73149994	rs7981432	T	A	.	intergenic variant	20.340	5.87e-06	0.777	0.066234	0.793	1.0e-06	0.78	(0.71,0.86)	0.88	0.0	2,283	384,969
chr13	75670983	chr13:75670983	TAC	!	*LMO7*	.	37.907	1.02e-06	0.799	0.44147	1.079	4.0e-05	0.84	(0.78,0.91)	0.0058	86.9	2,283	384,969
chr16	14968366	rs373248556	A	G	.	intergenic variant	0.326	6.39e-06	3.201	0.50975	0.018	7.4e-06	3.17	(1.91,5.25)	0.40	0.0	2,283	384,969
chr17	57827826	chr17:57827826	CAAA	!	.	.	17.612	1.74e-08	1.349	0.91155	1.014	1.7e-07	1.29	(1.17,1.42)	0.036	77.3	2,283	384,969
chr17	75938584	rs528106498	A	G	*FBF1*	upstream gene variant	0.002	2.74e-06	95.021	0.69588	0.018	3.6e-06	88.08	(13.24,586.08)	0.41	0.0	2,283	384,969
chr17	80334570	rs1279468131	G	A	*RNF213*	intron variant	0.002	4.12e-06	86.349	0.28318	0.017	2.7e-05	51.28	(8.15,322.52)	0.029	78.9	2,283	384,969
chr18	74246815	rs1465867509	T	C	.	intergenic variant	0.002	7.42e-06	80.214	0.49181	0.018	1.7e-05	63.79	(9.62,422.85)	0.16	50.3	2,283	384,969
chr21	34448298	rs142762112	A	AC	*KCNE1*	3 prime UTR variant	0.904	8.06e-07	2.266	0.33674	1.355	1.5e-06	2.03	(1.52,2.70)	0.15	51.8	2,283	384,969
Tics disorder (TD)																		
chr9	**136672271**	**rs564796941**	**C**	**CAAGA**	***EGFL7***	**frame-shift variant**	**0.328**	**4.52e-13**	**5.05**	**0.29**	**4.15**	**2.6e-13**	**5.02**	**(3.26,7.74)**	**0.89**	**0.0**	**1,721**	**393,872**
chr9	136671924	rs755785262	G	A	*EGFL7*	splice acceptor variant	0.075	0.113	2.84	.	.	0.113	2.84	(0.781,10.33)	.	.	1,322	326,339
chr9	136671015	rs746089480	A	G	*EGFL7*	splice donor variants	0.410	0.332	1.39	.	.	0.332	1.39	(0.715,2.704)	.	.	1,322	326,339

A large sample size is required to ascertain the association of p.Lys271ThrfsTer37-*EGFL7*. *EGFL7* is not a constrained gene (Lek et al., 2016). It encodes a 273 amino acid secreted protein (epidermal growth factor-like domain) and plays an important role in angiogenesis and cell trafficking (Usuba, Pauty, Soncin, & Matsunaga, 2019). A recent study has shown that *EGFL7* expression is increased in CNS vasculature of patients with multiple sclerosis (MS) and in mice with experimental autoimmune encephalomyelitis (EAE) where it may be used to reduce inflammation (Larochelle et al., 2018).

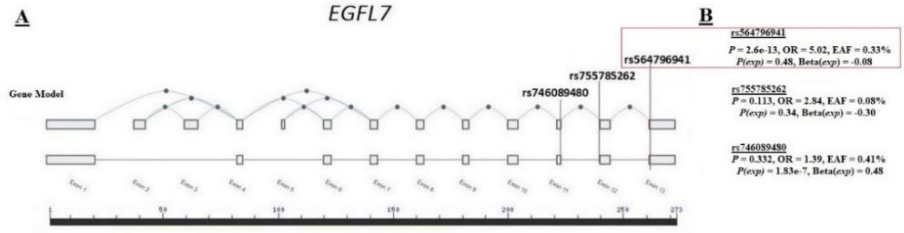


Figure 11: *EGFL7* model and loss-of–function sequence variants found in *EGFL7* with association data and RNA expression effect sizes. (A) *EGFL7* gene model with three loss-of-function variants, the frameshift variant (highlighted with red box, rs564796941-C) associates with TD. The variant is present in the last exon of *EGFL7*, while the other two loss-of-function variants are located in exon 10 and 11 of the gene. (B) Association summary statistics of three loss-of-function variants and their impact on RNA expression data, where *P* is the p-value for association of TD, and OR is the estimated odds ratio for TD, *P(exp)* is p-value for association of RNA expression, and Beta(*exp*) is standardized effect estimate on RNA expression.

4.4 GWAS meta-analysis of obsessive-compulsive disorder (unpublished data)

In this study, the largest GWAS meta-analysis of obsessive-compulsive disorder (cases = 8,317; controls = 1,060,098) was performed by combining GWAS summary data of OCD from Iceland, UKB, Norway, US, Denmark, Finland, and the Psychiatric genomics consortium (PGC). Therein, 15.8 million sequence variants were tested for association with OCD. The meta-analysis identified one missense variant associating with OCD: rs3733709 replacing isoleucine at position 289 with threonine in *PCDHA3* at 5q31.3 (OR = 0.866, $P = 2.3\times10^{-8}$, **Figure 12**). This association was not confirmed in an independent sample from Denmark (iPSYCH, OCD cases = 4,509, controls = 38,392, $P = 0.491$, OR = 1.012). The top associations for OCD are presented in **Table 5.**

The gene-based genome wide association analysis (through MAGMA) of OCD meta-analysis identified 12 significant genes located at 5q31.1 including the *PCDHA3* gene ($P = 6.7\times10^{-8}$, **Figure 12b**). Genes at other loci were not significant.

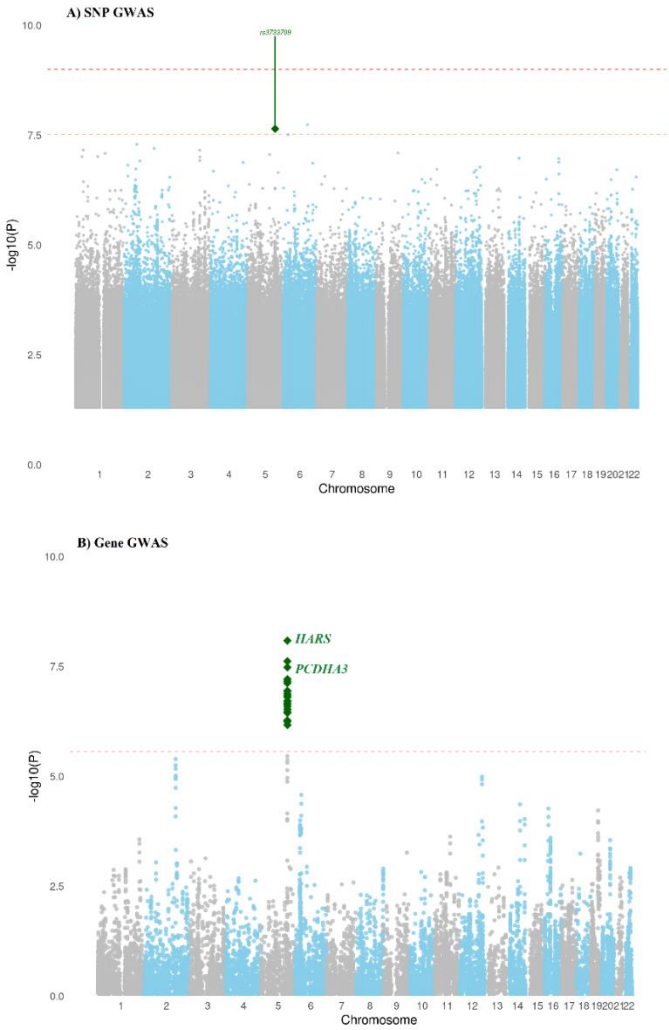


Figure 12: Manhattan plot showing meta-analysis of (A) SNP/Indel GWAS and (B) Gene-GWAS of OCD meta-analysis from Iceland, UKB, Norway, Denmark, US, Finland, and PGC cohorts (cases = 8,317; controls = 1,060,098). In 'A' the red dotted line represents SNP/Indels weighted threshold based on variant annotation (Sveinbjornsson et al., 2016) ($P < 0.05/4.29\times10^{-7} = 1.02\times10^{-9}$), orange line for missense variants ($P < 3.0\times10^{-8}$) and in 'B' the red line is the Bonferroni threshold for Gene-GWAS ($P < 0.05/18{,}540 = 2.7\times10^{-6}$). For SNP gwas plot, only nominally significant ($P < 0.05$) variants were used to generate Manhattan plots.

Table 5: Association results of top sequence variants from meta-analysis of OCD (cases = 8,317 controls = 1,060,098). Chr is for chromosome, **Pos_hg38** is position in hg38, **rsID** is the dbSNP ID of the variants, **EA** is effect allele, **OA** is other allele, **EAF%** is effect allele frequency in percentage, **Gene** is the closest genes within 500Kb, **category** is variant effect predictor annotation of effect EA, ***P*** is p-value for association, **OR** is estimated odds ratio for the effect allele, **Phet** is p-value for heterogeneity test, **I2** is percentage of variation across cohorts that is due to heterogeneity rather than by chance. Sample numbers from Iceland (cases = 2,286; controls = 137,961), Norway (cases = 145; controls = 91,416), PGC (psychiatric genomics consortium) (cases = 2,688; controls = 7,037), UKB (cases = 1,511; controls = 429,427), US (cases = 263; controls = 26,592), Finland (cases = 790; controls = 162,180), Denmark (cases = 634, controls = 91,416), and the combined meta-analysis (cases = 8,317; controls = 1,060,098).

Top variants from OCD Meta-analysis								Combined Meta-Analysis (Cases = 8,317; Controls = 1,060,098)			
Chr	**Pos**	**rsName**	**EA**	**OA**	**Gene**	**Category**	**EAF%**	***P***	**OR**	**Phet**	**I2**
chr1	54779866	rs1731	C	T	*TTC22*	3 prime UTR variant	1.746	2.2e-06	1.39598	0.262	21.994
chr1	63710853	s15078125	C	T	*PGM1*	regulatory region variant	0.674	4.62e-07	2.04071	0.428	0.000
chr1	224253241	rs5933214(	G	A	*NVL*	intron variant	20.899	8.8e-07	0.89395	0.616	0.000
chr11	83298357	rs612251	G	A	*CCDC90B*	intergenic variant	28.288	3.02e-06	0.91375	0.224	26.752
chr11	112964041	rs57715877	A	T	*NCAM1*	intron variant	31.989	1.24e-06	1.09527	0.961	0.000
chr12	23649076	s11779425	A	T	*SOX5*	intron variant	2.760	2.35e-06	1.27328	0.526	0.000
chr12	87095311	s77776468	C	T	*LOC105369878*	intergenic variant	0.026	2.88e-06	0.23707	0.783	0.000
chr12	105995093	rs2468203	A	G	*NUAK1*	intergenic variant	27.584	1.19e-06	0.91046	0.857	0.000
chr12	119685000	rs175878	T	C	*CIT,PRKAB1*	downstream gene variant	32.416	2.36e-06	0.91594	0.135	38.558
chr12	124615757	s19092593	A	G	*NCOR2*	intergenic variant	2.051	3.86e-06	0.71191	0.148	38.703
chr14	26065602	rs75210839	C	T	*NOVA1*	intergenic variant	0.000	4.87e-06	1.21300	0.377	6.244
chr14	76077030	rs2160880	A	G	*IFT43*	intron variant	47.579	3.06e-06	1.08210	0.841	0.000
chr16	78617776	rs9933350	A	C	*WWOX*	intron variant	12.635	4.56e-06	1.12176	0.642	0.000
chr16	85621329	rs72801186	A	G	*GSE1*	intron variant	5.318	1.21e-06	1.23788	0.251	24.437
chr19	1255720	rs57586121	G	A	*MIDN*	intron variant	0.768	2.54e-06	1.81194	0.709	0.000
chr19	40415372	s53254911	A	G	*PRX*	upstream gene variant	0.156	3.83e-06	0.49047	0.528	0.000
chr2	37149171	rs2307466	C	G	*EIF2AK2*	5 prime UTR variant	6.360	2.2e-06	1.18376	0.196	30.395
chr2	64665453	rs72814059	A	G	*SERTAD2*	intergenic variant	8.156	3.85e-06	1.14981	0.914	0.000
chr2	102875498	s20016246	T	C	*LINC01796*	intergenic variant	0.534	3.6e-06	1.58202	0.411	0.000
chr2	155500518	:144235668	G	C	*LINC01876*	intergenic variant	0.275	3.7e-06	0.20127	0.16	39.232
chr2	160518427	s55996739	C	G	*RBMS1*	intergenic variant	0.032	1.31e-06	2.89302	0.943	0.000
chr2	169668573	rs11686105	C	A	*CCDC173*	intron variant	44.083	2.04e-06	0.92182	0.87	0.000
chr2	225642884	rs13424677	A	T	*NYAP2*	intron variant	19.682	5.62e-07	0.89557	0.451	0.000
chr21	18436543	s57402706	C	A	*TMPRSS15*	intergenic variant	0.216	3.56e-06	2.25196	0.394	0.000
chr22	49553711	s56668531	T	G	*C22orf34*	intergenic variant	0.084	3.46e-06	3.34846	0.489	0.000
chr3	2599096	s55159228	T	G	*CNTN4*	intron variant	0.006	2.12e-06	2.87974	0.983	0.000
chr3	71616968	rs75363374	A	G	*FOXP1*	intergenic variant	2.255	3.24e-06	1.36697	0.658	0.000
chr3	78327849	s11309971	G	A	*ROBO1*	intergenic variant	3.850	3.82e-06	1.25383	0.636	0.000
chr4	27904637	rs6825286	T	G	*LINC02261*	intergenic variant	16.389	4.69e-06	1.11371	0.668	0.000
chr5	120603733	rs10061078	T	C	*PRR16*	intron variant	7.645	3.78e-06	1.15917	0.743	0.000
chr5	**140802063**	**rs3733709**	**C**	**T**	***PCDHA3***	**p.Ile289Thr:NP 061729.1**	**12.615**	**2.27e-08**	**0.86641**	**0.0946**	**44.474**
chr5	152629875	rs2964260	T	C	*LINC01470*	intergenic variant	46.501	4.65e-06	1.08058	0.178	32.681
chr6	21283307	s76806318	T	C	*CDKAL1*	intergenic variant	0.017	1.63e-06	0.01450	1	0.000
chr6	27692729	rs9380007	C	T	*LINC01012*	intergenic variant	36.848	3.1e-06	0.92035	0.734	0.000
chr6	60644221	s18198228	C	T	*MTRNR2L9*	intergenic variant	0.434	4.48e-06	1.64576	0.506	0.000
chr6	112689714	rs1814561	G	A	*LOC105377949*	intergenic variant	0.636	2.5e-06	0.89655	0.943	0.000
chr6	127931780	s87946898	A	T	*THEMIS*	intergenic variant	0.195	4.5e-06	0.26751	0.543	0.000
chr6	150332573	s13890657	T	G	*IYD*	intergenic variant	1.186	1.91e-06	1.47432	0.731	0.000
chr7	124318842	s19230202	T	G	*LOC107986841*	intergenic variant	0.279	2.67e-06	2.39289	0.629	0.000

4.5 GWAS meta-analysis of Restless legs syndrome (paper IV)

Paper IV reports the largest, to date, GWAS meta-analysis of restless legs syndrome (RLS) including 10,257 cases, and 470,725 controls from five populations (Iceland, UK, Denmark, US, and the Netherland).

RLS is a complex sensorimotor disorder with a prevalence ranging from 5 to 18.8% in European populations (M. Didriksen et al., 2017). Symptoms include distressing sensations in the extremities and overwhelming urge to move the legs. These symptoms intensify when sitting or lying down. The disorder can cause reduced quality of life and sleep and impair cognition and mental well-being. Despite the high prevalence and serious health impact of the disorder, there are currently no adequate treatments for RLS as available drugs target symptoms and are fraught with side effects. Drug discovery and development may be hampered by an incomplete understanding of the pathophysiology of RLS.

4.5.1 Novel variants associated with RLS

In this GWAS meta-analysis, 15.8 million DNA sequence variants were tested for association with RLS using 505,959 individuals of European ancestry (a discovery sample of 10,257 cases, and 470,725 controls and a replication sample of 6,651 cases and 18,326 controls). In the discovery meta-analysis, 19 of the 20 previously reported RLS associated sequence variants were confirmed, and three novel associations with RLS were found and replicated in the combined, independent sample (**Figure 13, Table 6**). The novel RLS associated sequence variants are; rs10068599-T in an intron of *RANBP17* on 5q35.1 (OR = 1.09, $P = 6.9\times10^{-10}$, 95% CI: 1.06–1.12), rs112716420-G in close proximity of *MICALL2* on 7p22.3 (OR 1.25, $P = 1.5\times10^{-18}$, 95% CI: 1.19–1.31) and rs10769894-A near *LMO1* and *STK33* on 11p15.4 (OR = 0.90, $P = 9.4\times10^{-14}$, 95% CI: 0.88–0.93) (**Table 6**).

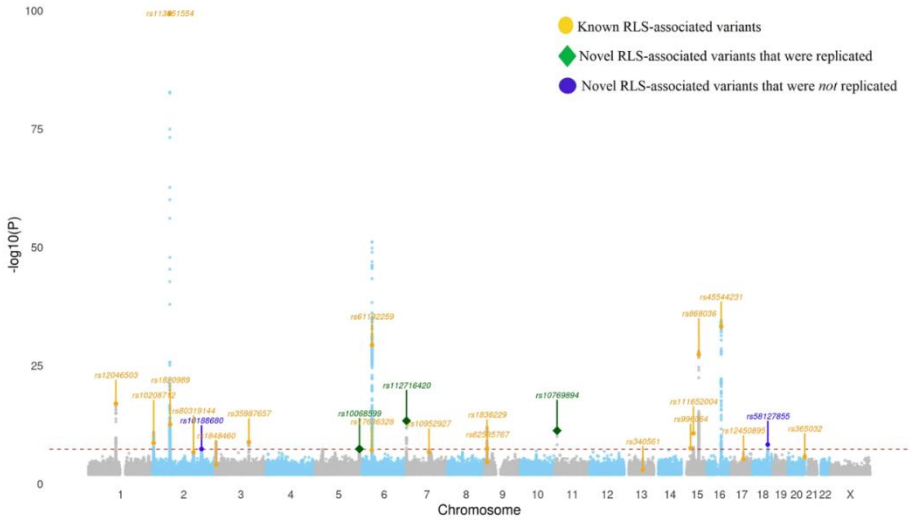


Figure 13: Manhattan plot displaying results from the RLS discovery meta-analysis for *N* = 480,982 independent biological samples. Variants labelled orange were previously reported variants. Variants labelled blue and green are novel variants (five) that were tested in a follow-up sample. Of the five novel variants, three were confirmed (green diamond shape) in the follow up analysis and met the genome-wide significance threshold, whereas two did not (Table 7).

Table 6: Sequence variants associated with RLS. EA is effect allele, **OA** is other allele, and **EAF** is effect allele frequency, **OR** is estimated odds ratio of the effect allele, ***P*** refers to association P-value of the tested allele, **Gene** is the closest gene within 500Kb. [a] is for discovery meta-analysis using 10,257 cases, and 470,725 controls, [b] Follow up analysis of top five signals was carried out in two independent replication samples: EU-RLS-GENE cohort (cases/controls = 6,228/10,992) and the RBC-Omics cohort (423/7,334) (See online Supplementary Table 1 for details and Supplementary Table 2, which displays results for all known RLS-associated variants). [c] the combined analysis comprises both the discovery sample as well as the two replication samples. * Represents significant P-value for replication samples after correcting for multiple testing: $P < 0.05/5/2 = 0.005$. [d] Reference: PMID: 29029846, [e] Combined meta-analysis of published GWAS (PMID: 29029846) and data from this study.

Novel variants associated with RLS						Discovery[a] Cases = 10,257 Controls = 470,725		Follow up analysis[b] Cases = 6,651 Controls = 18,326		Combined analysis[c] Cases = 16,908 Controls = 489,051	
rsID	**Chr**	**Position (hg38)**	**EA/OA**	**EAF**	**Gene**	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***
rs10188680	Chr2	189,584,800	T/A	0.4	*SLC40A1*	1.09 (1.06-1.13)	4.3×10^{-08}	1.04 (0.99-1.09)	0.13	1.07 (1.05 - 1.11)	5.4×10^{-08}
rs10068599	chr5	171,001,975	T/C	0.3	*RANBP17*	1.1 (1.06-1.13)	4.3×10^{-08}	1.07 (1.03-1.11)	0.0031*	1.09 (1.06-1.12)	6.9×10^{-10}
rs112716420	chr7	1,343,010	G/C	0.1	*MICALL2/UNCX*	1.24 (1.18-1.30)	4.9×10^{-14}	1.27 (1.17-1.37)	5.6×10^{-06}*	1.25 (1.19-1.31)	1.5×10^{-18}
rs10769894	chr11	8,313,948	A/G	0.5	*LMO1*	0.89 (0.86-0.93)	5.8×10^{-12}	0.92 (0.87-0.97)	0.0029*	0.9 (0.88-0.93)	9.4×10^{-14}
rs58127855	Chr18	59,943,413	T/C	0	*PMAIP1*	4.72 (4.20-5.24)	5.1×10^{-09}	0.91 (-0.01-1.83)	0.84	3.03 (2.01 - 4.97)	6.3×10^{-07}
Known variants associated with RLS[d]						**Current study** Cases = 10,257 Controls = 470,725		**Literature[d]** Cases = 15,126 Controls = 95,725		**Combined analysis[e]** Cases = 25,383 Controls = 566,450	
rsID	**Chr**	**Position (hg38)**	**EA/OA**	**EAF**	**Gene**	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***
rs10208712	chr2	3986856	G/A	0.4	.	0.91 (0.88-0.94)	2.34×10^{-09}	0.9 (0.87-0.93)	3.78×10^{-15}	0.9 (0.88-0.92)	5.9×10^{-23}
rs10952927	chr7	88729746	G/A	0.1	.	1.13 (1.09-1.17)	1.9×10^{-09}	1.17 (1.13-1.21)	1.86×10^{-15}	1.15 (1.12-1.18)	4.1×10^{-21}
rs111652004	chr15	47068169	T/G	0.1	.	0.83 (0.77-0.88)	2.2×10^{-11}	0.84 (0.79-0.89)	1.05×10^{-10}	0.83 (0.79-0.87)	1.5×10^{-20}
rs113851554	chr2	66523432	T/G	0.1	*MEIS1*	1.89 (1.83-1.94)	4.5×10^{-100}	2.16 (2.11-2.21)	1.1×10^{-180}	2.03 (1.99-2.07)	3.3×10^{-276}
rs12046503	chr1	106652717	C/T	0.4	.	1.15 (1.11-1.18)	1.09×10^{-17}	1.18 (1.15-1.20)	3.32×10^{-32}	1.16 (1.14-1.18)	7.1×10^{-48}
rs12450895	chr17	48695414	A/G	0.2	.	1.09 (1.05-1.13)	5.69×10^{-06}	1.09 (1.06-1.12)	4.87×10^{-08}	1.09 (1.07-1.11)	1.3×10^{-12}
rs12962305	chr18	44290278	T/C	0.3	.	1.03 (1.01-1.05)	0.0113	1.11 (1.08-1.14)	1.37×10^{-10}	1.06 (1.04-1.08)	4.5×10^{-09}
rs17636328	chr6	37522755	G/A	0.2	.	0.9 (0.86-0.94)	7.63×10^{-08}	0.89 (0.86-0.92)	6.43×10^{-11}	0.89 (0.86-0.92)	2.7×10^{-17}
rs1820989	chr2	67842758	A/C	0.5	.	1.12 (1.09-1.15)	2.86×10^{-13}	1.14 (1.11-1.16)	1.23×10^{-20}	1.13 (1.11-1.15)	3.1×10^{-32}
rs1836229	chr9	8820573	G/A	0.5	*PTPRD*	0.92 (0.89-0.95)	3.68×10^{-08}	0.9 (0.87-0.93)	1.94×10^{-15}	0.91 (0.89-0.93)	6.2×10^{-22}
rs1848460	chr3	3406460	T/A	0.3	.	1.06 (1.03-1.08)	7.3×10^{-05}	1.13 (1.10-1.16)	5.38×10^{-14}	1.09 (1.07-1.11)	3.0×10^{-15}
rs340561	chr13	72274018	T/G	0.2	.	1.07 (1.03-1.10)	0.001	1.09 (1.06-1.12)	3.93×10^{-08}	1.08 (1.06-1.10)	2.5×10^{-10}
rs35987657	chr3	130816723	G/A	0.3	.	0.9 (0.87-0.94)	1.45×10^{-09}	0.9 (0.87-0.93)	4.37×10^{-13}	0.9 (0.88-0.92)	3.9×10^{-21}
rs365032	chr20	64164052	G/A	0.3	*MYT1*	1.09 (1.05-1.12)	2.13×10^{-06}	1.13 (1.10-1.16)	3.36×10^{-14}	1.11 (1.09-1.13)	1.5×10^{-18}
rs45544231	chr16	52598818	G/C	0.4	.	0.82 (0.79-0.85)	5.71×10^{-34}	0.81 (0.78-0.84)	4.72×10^{-48}	0.81 (0.79-0.83)	3.9×10^{-80}
rs61192259	chr6	38486186	C/A	0.4	*BTBD9*	0.83 (0.80-0.86)	4.71×10^{-30}	0.76 (0.73-0.79)	1.36×10^{-78}	0.79 (0.77-0.81)	1.9×10^{-103}
rs62535767	chr9	9290311	T/C	0.3	*PTPRD*	0.93 (0.89-0.96)	2.2×10^{-05}	0.91 (0.88-0.94)	3.13×10^{-10}	0.92 (0.89-0.94)	4.8×10^{-14}
rs80319144	chr2	158343323	T/C	0.2	*CCDC148*	0.91 (0.88-0.95)	2.11×10^{-07}	0.89 (0.86-0.92)	3.18×10^{-14}	0.9 (0.88-0.92)	5.5×10^{-20}
rs868036	chr15	67762675	T/A	0.3	*MAP2K5*	0.83 (0.79-0.86)	4.67×10^{-28}	0.8 (0.77-0.83)	1.09×10^{-48}	0.81 (0.79-0.83)	1.8×10^{-74}
rs996064	chr15	35916797	T/A	0.1	.	1.21 (1.14-1.27)	2.8×10^{-08}	1.21 (1.15-1.27)	2.96×10^{-09}	1.21 (1.16-1.26)	4.4×10^{-16}

4.5.2 Cis-colocalization analysis of RLS variants

To investigate whether the RLS variants exert their impact through gene expression, we performed a cis-colocalization analysis of RLS variants using 49 tissues reported in the GTEx database, i.e., whether any of the RLS variants are also the top eQTL signals of the respective gene and tissue.

Therein, the cis-eQTL data for 11 sequence variants impacting 17 genes was found. Of the 11 variants with data, 10 strongly associate with cis-gene expression ($P < 3.3\times10^{-6}$). Six of these 10 variants are in LD ($r^2 > 0.3$) with top-

eQTL for the respective gene. To ascertain that RLS variants and top-eQTLs share the same signal, a two-way approximate conditional analysis implemented in COJO (J. Yang et al., 2012) was employed. Therein, conditional analysis using RLS effect sizes showed that four RLS variants and eQTLs share the same signal (**Figure 14**). Additionally, conditional analysis using GTEx effect sizes also confirmed the same associated signals. Hence, four RLS variants (rs10068599-T, rs1063756-CACAG, rs12450895-A, and rs3784709-T) co-localize with top eQTLs for five genes respectively (*RANBP17, CASC16, HOXB2, MAP2K5*, and *SKOR1*) (**Figure 14)**.

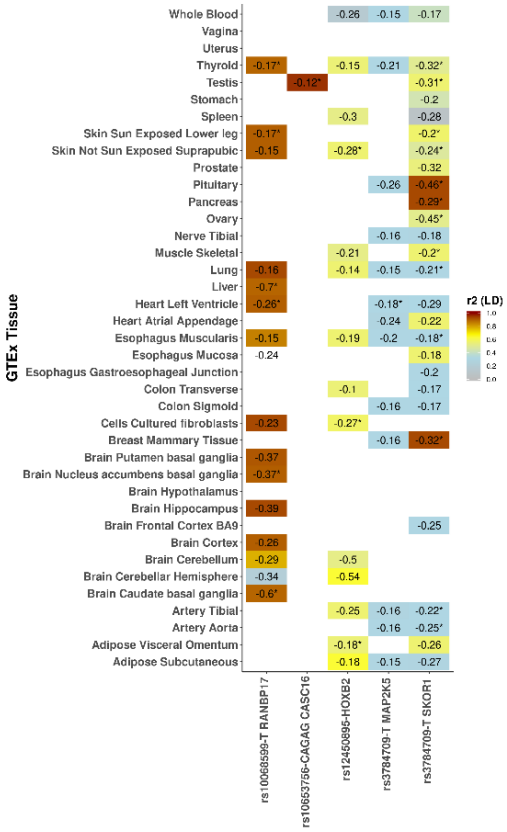


Figure 14: Cis co-localization of RLS variants using GTEx eQTLs data. Displaying eQTL variants for RLS variants that significantly associate with cis-gene expression at least in one tissue tested are in linkage disequilibrium (LD) ($r2 > 0.30$) and share the same causal signal (as confirmed through approximate conditional analysis) with the top eQTL variant of the respective genes. Cis-eQTL effect estimates (normalized) are provided and those sharing same causal signal (COJO conditional analysis) with eQTL and are Bonferroni significant ($P < 3.3\times10^{-6}$) are labelled with an asterisk.

4.5.3 Shared genetic architecture between RLS and life-style traits

A recent study has shown that RLS is associated with lower educational attainment, life-style and cognitive traits (cognitive performance, educational attainment, neuroticism score, smoking behaviour, and percentage of fat in the legs and in the whole body) (Barbara Schormair et al., 2017). The RLS polygenic risk score (RLS-PRS) was used to perform phenome-wide association analysis of 12,075 binary and continuous traits ($P_{threshold} < 0.05/12{,}075 = 4.14\times10^{-6}$).
Our analysis confirmed the prior findings that higher RLS-PRS burden is negatively associated with educational attainment ($P = 2.7\times10^{-25}$, β = -0.02, S.E. = 0.002), cognitive performance ($P = 4.4\times10^{-7}$, β = -0.01, S.E. = 0.002), and age at first time giving birth ($P = 5.9\times10^{-16,}$ β = -0.02, S.E. = 0.003), and increased risk of smoking behavior ($P = 1.39\times10^{-6}$, OR = 1.05, S.E. = 0.009). The PRS score furthermore associates positively with neuroticism ($P = 8.0\times10^{-23}$, β = 0.01, S.E. = 0.002), as well as fat percentage in legs ($P = 1.4\times10^{-10}$, β = 0.01, S.E. = 0.002), and in the whole body ($P = 4.7\times10^{-7}$, β = 0.008, S.E. = 0.002) (**Table 7**). Furthermore, genomewide genetic correlation analysis through LD score regression (B. K. Bulik-Sullivan et al., 2015) confirmed association results from PRS analysis (**Table 7**).

Table 7: Displaying results from the association of RLS-PRS with several binary health-related traits and their genetic correlation with RLS GWAS meta-analysis. [a] Refers to the association results from RLS-PRS predictions, [b] is for the summary statistics from the genetic correlation analysis using LD score regression (B. K. Bulik-Sullivan et al., 2015). Cases is for the number of cases, and controls is for the number of controls used for the analysis, *N* is for the total sample size, *P* is p-value for the association, OR refers to odds ratio predicted by RLS-PRS, R2 is phenotypic variance explained by RLS-PRS, *rg* is for genetic correlation, S.E. is the standard error for the GC analysis, and Z is the z-score estimate from the GC analysis.

Phenotype				PRS analysis[a]				GC analyis[b]			
Phenotype name	**Cases**	**Controls**	***N***	**OR**	**Beta**	***P***	**R2**	**rg**	**S.E**	**Z**	***P***
Gastro-intestine tract	77,433	331,211	408,644	1.028	*NA*	6.27e-12	0.0183	0.3858	0.0839	4.5998	4.2292e-06
Upper gastrointestinal tract examination	45,348	363,304	408,652	1.034	*NA*	1.99e-11	0.0217	0.4206	0.0932	4.5121	6.4187e-06
Gastritis and duodenitis	28,747	379,817	408,564	1.035	*NA*	2.95e-8	0.0187	0.4362	0.1016	4.2953	1.7443e-05
Diaphragmatic hernia	26,926	381,726	408,652	1.036	*NA*	3.67e-8	0.0191	0.3148	0.0807	3.9001	9.615e-05
Extrapyramidal movement disorders	3,737	408,819	412,556	1.238	*NA*	5.37e-38	0.4104	0.944	0.2467	3.8266	0.0001
Duodenum procedue	41,365	367,279	408,644	1.030	*NA*	9.89e-9	0.0166	0.4182	0.1119	3.7379	0.0002
Gastro Esophageal reflux disease	23,050	385,602	408,652	1.0361	*NA*	1.939e-07	0.01873	0.3587	0.1012	3.545	0.0004
Spondylosis	7,642	401,002	408,644	1.0569	*NA*	1.987e-06	0.03246	0.5377	0.1523	3.5315	0.0004
Spine procedure	12,439	396,213	408,652	1.0448	*NA*	1.585e-06	0.02361	0.3392	0.099	3.4274	0.0006
Stomach Procedue	35,002	373,642	408,644	1.0271	*NA*	2.155e-06	0.01228	0.342	0.1004	3.4055	0.0007
Primary hypertension	77,566	331,086	408,652	1.0218	*NA*	2.122e-07	0.00982	0.207	0.0682	3.0332	0.0024
Pure hypercholesterolaemia	33,184	375,468	408,652	1.0278	*NA*	2.899e-06	0.01193	0.2385	0.0789	3.0238	0.0025
Spondylosis	4,640	404,004	408,644	1.0726	*NA*	2.461e-06	0.04632	0.4953	0.1689	2.9322	0.0034
Asthma	25,929	382,715	408,644	1.0322	*NA*	8.469e-07	0.0157	0.1967	0.0825	2.3845	0.0171
Tobacco use	11,901	396,743	408,644	1.0461	*NA*	1.397e-06	0.02451	0.2044	0.0864	2.3667	0.0179
Education Years	.	.	405282	*NA*	-0.0155	2.84e-23	0.02407	-0.1731	0.0449	-3.8523	0.0001
Fluid Intelligence	.	.	204073	*NA*	-0.01044	2.53e-06	0.010845	-0.0891	0.0503	-2.0176	0.0253
Fat percentage in Legs	.	.	401750	*NA*	0.01013	1.43e-10	0.010234	0.1368	0.0415	3.3004	0.001
Whole body fat	.	.	401061	*NA*	0.0077	1.26e-06	0.0058546	0.1028	0.0404	2.5451	0.0109
Neuroticism	.	.	332083	*NA*	0.0145	8.03e-17	0.020856	0.1729	0.0819	2.112	0.0347

4.6 Cross disorder genetic analysis identifies shared genetic effect with insomnia (unpublished data)

Our analysis detected that childhood neuropsychiatric and involuntary movement disorders share some genetic architecture (**Figure 1**). However, it remains unclear how this genetic overlap is shaped and whether some sequence variants have pleiotropic effect on correlated traits impacting through shared biological pathway(s). To understand this complex genetic interplay, the phenome-wide genetic correlation analysis of five disorders from this study vs. 1,140 published GWAS studies was performed (**Figure 15**). For the common genetically correlated traits, the hierarchical clustering method was used to identify the latent correlated components. Additionally, to understand the causal relationship of genetically correlated traits, the genome-wide significant markers were used as instrumental variables (IVs). Therein, a two-sample Mendelian randomization approach was employed by using effect estimates of correlated traits as exposure phenotypes and the effect estimates of IVs for neuropsychiatric disorders as the outcome traits.

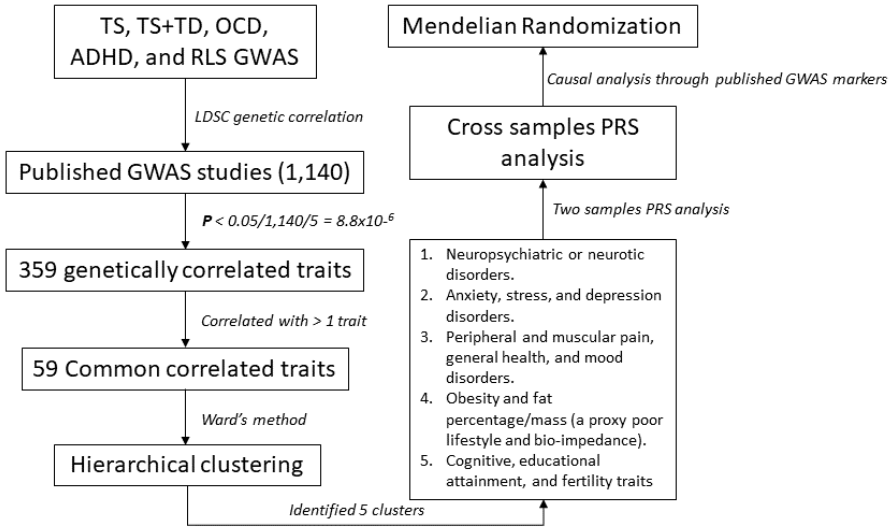


Figure 15: Study scheme for cross disorder genetic analysis.

4.6.1 Phenome-wide genetic correlation analysis

We regressed GWAS results of childhood neuropsychiatric and involuntary movement disorders from this study (ADHD, OCD, TS, TS+TD, and RLS) against 1,140 published genome-wide association studies with effective sample size of

over 5,000 employing LD score regression ($P < 0.05/1,140/5 = 8.8\times10^{-6}$). The LD score regression analysis detected correlations of these disorders with hundreds of traits and diseases, including neuropsychiatric disorders, neurological diseases, brain structures, cognitive traits, sleep disorders, substance use disorders, mood and personality disorders, behavioral phenotypes, anthropometric traits, bio-impedance measures (as proxy for body-fat percentage), musculoskeletal and neuropathic pain, and life-style traits (**Figure 16 to Figure 19**). Among the childhood neuropsychiatric disorders, ADHD shows broad genetic correlation (with 356 of 1,140 tested traits) with phenotypes from 25 categories (**Figure 16**). Compared to ADHD, OCD has weaker genetic correlation with the tested traits and fewer significant correlations (95 of the 1,140 tested traits) (**Figure 17**).

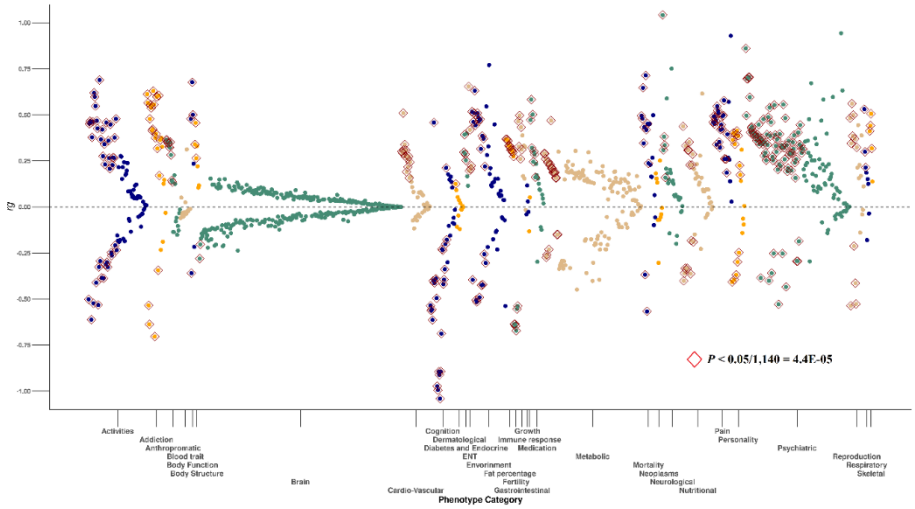


Figure 16: Phenome-wide genetic correlation between ADHD and 1,140 published GWAS studies. Each dot is an estimate of genetic correlation (*rg*) between ADHD and one of the tested traits (binned into a phenotype category), where x-axis represents phenotype (category) and y-axis showing its genetic correlation (*rg*). The significant associations ($P < 0.05/1,140/5 = 8.8\times10^{-6}$) are highlighted with red-diamond shape.

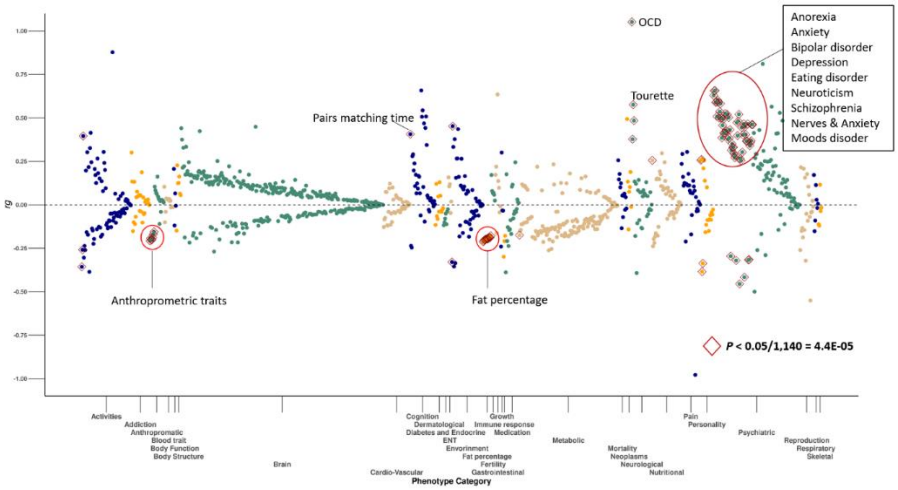


Figure 17: Phenome-wide genetic correlation between OCD and 1,140 published GWAS studies. Each dot is an estimate of genetic correlation (*rg*) between OCD and one of the tested traits (binned into a phenotype category), where x-axis represents phenotype (category) and y-axis showing its genetic correlation (*rg*). The significant associations ($P < 0.05/1,140/5 = 8.8\times10^{-6}$) are highlighted with red-diamond shape.

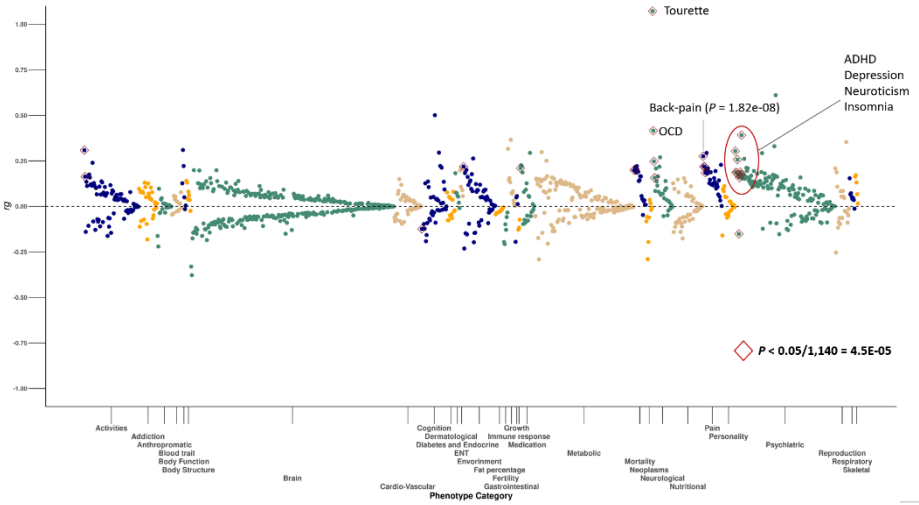


Figure 18: Phenome-wide genetic correlation between TS and 1,140 published GWAS studies. Each dot is an estimate of genetic correlation (*rg*) between TS and one of the tested traits (binned into a phenotype category), where x-axis represents phenotype (category) and y-axis showing its genetic correlation (*rg*). The significant associations ($P < 0.05/1,140/5 = 8.8\times10^{-6}$) are highlighted with red-diamond shape.

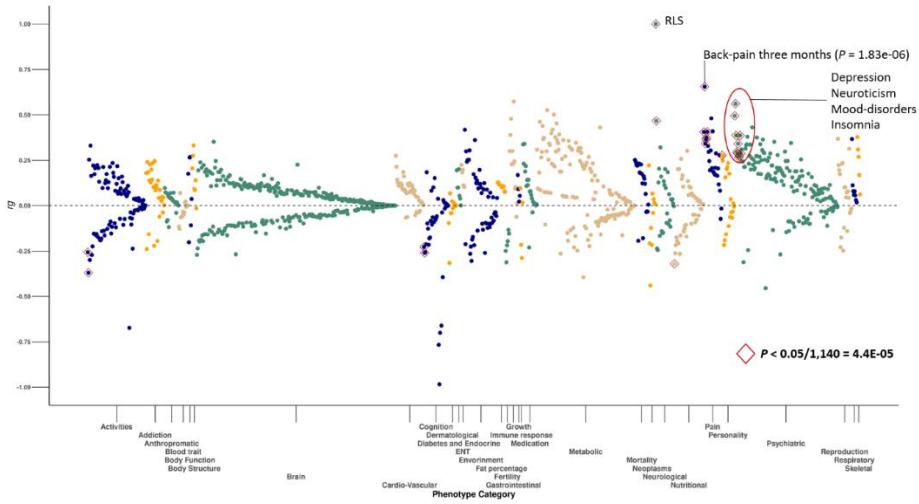


Figure 19: Phenome-wide genetic correlation between RLS and 1,140 published GWAS studies. Each dot is an estimate of genetic correlation (*rg*) between RLS and one of the tested traits (binned into a phenotype category), where x-axis represents phenotype (category) and y-axis showing its genetic correlation (*rg*). The significant associations ($P < 0.05/1{,}140/5 = 8.8\times10^{-6}$) are highlighted with red-diamond shape.

4.6.2 Hierarchical clustering of correlated traits

In the presence of a large set of genetically correlated traits, it is difficult to untangle the true genetic overlap of childhood neuropsychiatric disorders with involuntary movement disorder. To better understand this relationship, the hierarchical clustering approach was employed to identify clusters with shared genetic components among these traits. Subsequently, the common genetically correlated traits (at-least correlated with two traits and with $P < 0.05/1{,}140/5 = 8.8\times10^{-6}$) from pair-wise phenome-wide genetic correlation analysis of neuropsychiatric and involuntary movement disorders were selected. The analysis identified 59 such traits, diseases, or disorders.

Hierarchical clustering (Ward's method) of correlation estimates ('*rg*' from genetic correlation analysis through LD score regression (B. K. Bulik-Sullivan et al., 2015) identified that TS, and TS+TD form a cluster with RLS, which then form cluster with OCD and ADHD (**Figure 20**). Moreover, the hierarchical clustering analysis of these traits with 59 common traits identified five latent clusters (**Figure 20**). These clusters could be categorized into (1) neuropsychiatric or neurotic disorders, (2) anxiety, stress, or depression disorders, (3) clinical phenotypes with peripheral and muscular pain, general health, and mood disorders, (4) obesity and fat percentage/mass (a proxy for poor lifestyle and

bio-impedance), and (5) cognition, or learning/educational attainment, and fertility traits (**Figure 20**).

Among these the role of major depressive disorder (MDD), behavioral traits (smoking and neuroticism), obesity, and pain phenotypes are noticeable. MDD has been shown to positively correlate (genetically) with ADHD, TS, OCD, and RLS (Anttila et al., 2018; Didriksen et al., 2020). Moreover, smoking behavior has also been shown to positively correlate (genetically) with ADHD, and RLS (Anttila et al., 2018; Didriksen et al., 2020). Smoking is a marker of poor lifestyle in general, and as such associates with various important socio-behavioral phenotypes, such as socioeconomic status and educational attainment. Educational attainment has also been shown to negatively correlate with ADHD, TS, and RLS (Anttila et al., 2018; Didriksen et al., 2020).

Most strikingly, for obesity related traits, an interesting relationship between ADHD and OCD disorders was observed. Therein, ADHD positively correlates with obesity traits whereas OCD has negative correlation with obesity traits (**Figure 20**). This genetic correlation may partially be explained by their correlation with anorexia (**Figure 20**). ADHD correlates positively with obesity and negatively with eating disorder and OCD. OCD correlates positively with eating disorder and negatively with obesity. These finding further help understand indirect association.

Both ADHD and OCD both correlate positively with neuroticism, sch, bipolar, mdd and insomnia. This perhaps suggest that while they are on "opposite" ends of the impulsive (ADHD) to compulsive (OCD) spectrum, they both share genetics with psychiatric disorders.

The genetic correlation of the 3rd cluster involving clinical phenotypes with peripheral pain and muscular pain has not been explored much. Anttila et.al., has shown some evidence for genetic correlation of migraine with ADHD, and TS (Anttila et al., 2018). However, in our study, this correlation is not significant for TS after correcting for multiple testing ($rg_{\text{TS-Migraine}} = 0.15$, $P = 0.009$, $rg_{\text{ADHD-Migraine}} = 0.26$, $P = 2.45\times10^{-7}$). The shared genetic component of pain conditions (skeletal pain in back-hip-joints, and muscular pain in neck-shoulder and legs) with TS, TS+TD, RLS, and ADHD ($P < 0.05/1{,}140/5 = 8.8\times10^{-6}$) was identified. This is an interesting finding providing opportunities for further research.

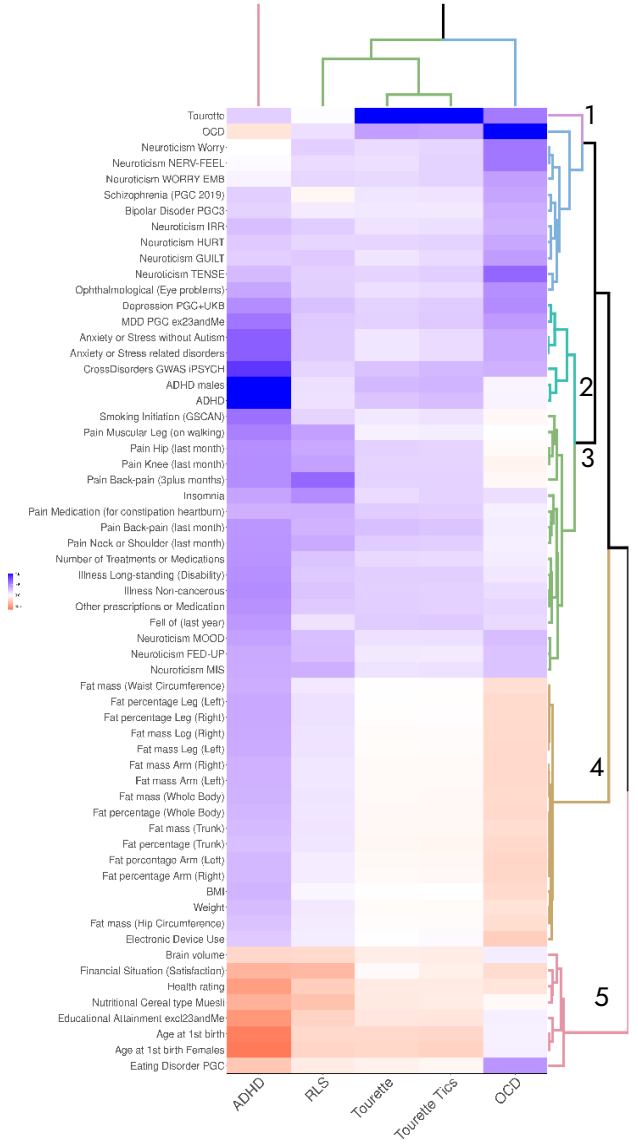


Figure 20: Genetic correlation and hierarchical clustering of common genetically correlated traits (at least with two of the five tested neuropsychiatric or involuntary movement disorders with $P < 0.05/1,140/5 = 8.8\times10^{-6}$). The hierarchical clustering (using Ward 's method) identify that TS, TS+TD, and RLS share common genetic structure that form cluster with OCD, and then form cluster with ADHD. Moreover, the hierarchical clustering of these neurological disorders identified five latent clusters that correlate with these disorders (1- neuropsychiatric or neurotic disorders; 2- anxiety, stress, and depression disorders; 3- clinical phenotypes with peripheral and muscular pain, general health and mood disorders; 4- obesity and fat percentage/mass (a proxy for bio-impedance); and 5- cognitive traits, educational attainment, and fertility traits).

4.6.3 Causal analysis (Mendelian randomization)

To disentanlge the causal effects between phenotypes, a two-sample Mendelian randomization (MR) approach can be employed using instrumental variables from common genetically correlated traits. The instrumental variables (IVs) used for MR analysis are sensitive to sample size and strength of association with the predictor phenotype (Morrison, Knoblauch, Marcus, Stephens, & He, 2020). The robustly associated GWAS significant variants and their effect sizes from the largest available studies of smoking behavior (Mengzhen Liu et al., 2019; Xu et al., 2020), depression (Howard et al., 2019), BMI (Yengo et al., 2018), neuroticism (Mats Nagel et al., 2018), insomnia (Kyoko Watanabe et al., 2020), ADHD (Ditte Demontis et al., 2019), schizophrenia (Consortium, 2014), bipolar disorder (Eli A Stahl et al., 2019), intracranial volume (Philip R Jansen et al., 2019), subcortical brain structures (Satizabal et al., 2019), intelligence (Savage et al., 2018), and educational attainment (J. J. Lee et al., 2018b) were used as instrumental variables (IVs). Therein, 90 independent tests were performed using IVs and effect sizes from these studies as exposure phenotype compared to their effect sizes in childhood neuropsychiatric and involuntary disorders as outcome phenotype ($P < 0.05/18/5 = 5.5\times10^{-4}$).

Amongst the tested IVs from 18 genetic studies, the insomnia associated sequence variants exert strongest impact on all the tested childhood neuropsychiatric, and involuntary movement disorders ($\beta > 0.85$, $P < 6.7\times10^{-10}$, **Figure 21**). It implies that insomnia has causal effect on the neurological disorders. Notably, a strong genome-wide genetic correlation between insomnia and neurological disorders was also observed (**Figure 16 - 19**). The genetic correlation analysis is in line with the reported finding that polygenic risk score of neurodevelopmental disorders correlates with sleep disorders (Ohi et al., 2021). Future studies are required to understand whether there is a bi-directional effect between neurological disorders and insomnia (sleep disturbances).

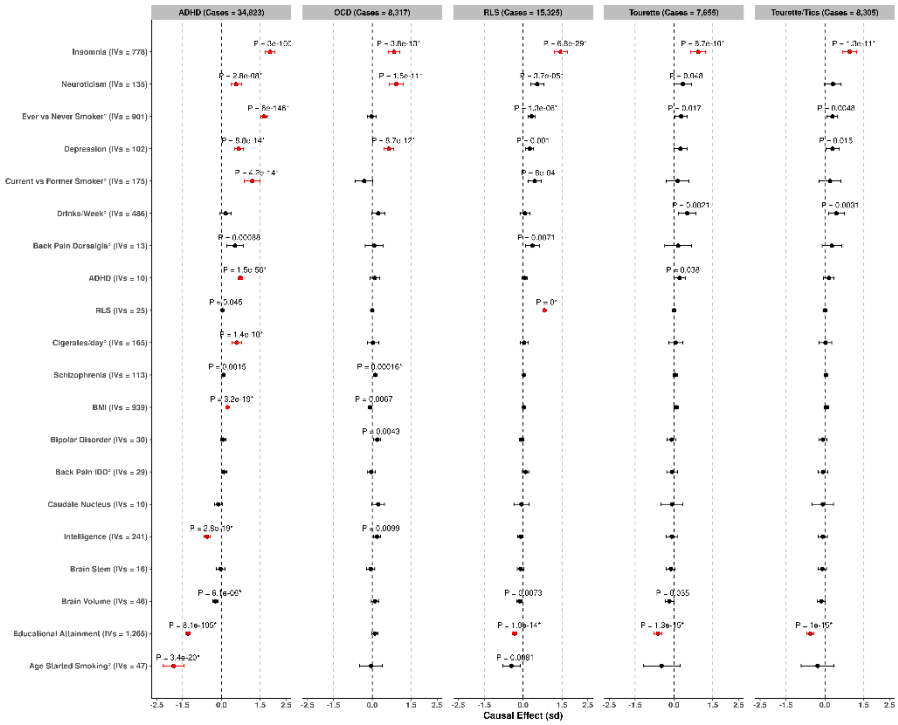


Figure 21: Causal analysis of GWAS significant markers as IV from common genetically correlated traits. The red dotted vertical line is set at '0' represents no effect.

Moreover, top GWAS significant sequence variants associated with smoking behavior, neuroticism, depression, body-mass-index, intelligence, educational attainment, and ICV exert large effects on ADHD (β = 1.65, $P = 6.0\times10^{-146}$, β = 0.57, $P = 2.8\times10^{-8}$, β = 0.67, $P = 8.8\times10^{-14}$, β = 0.53, $P = 1.4\times10^{-42}$, β = -0.57, $P = 2.8\times10^{-19}$, β = -1.29, $P = 2.8\times10^{-195}$, β = -0.23, $P = 2.8\times10^{-6}$, respectively). These findings are in line with the genetic correlation of ADHD and these traits previously reported (Anttila et al., 2018; Klein et al., 2019; Vink, Treur, Pasman, & Schellekens, 2020; F. Zhang et al., 2021).

Furthermore, the MR analysis highlighted that sequence variants associated with insomnia, neuroticism, depression, and schizophrenia show causal relationship with OCD (β = 0.85, $P = 3.8\times10^{-13}$, β = 0.93, $P = 1.5\times10^{-11}$, β = 0.65, $P = 8.7\times10^{-12}$, β = 0.11, $P = 1.6\times10^{-4}$, respectively). While genetic correlation between OCD and depression, neuroticism, and schizophrenia has been reported before (Anttila et al., 2018).To the best of our knowledge, the genetic correlation of OCD with insomnia is a novel observation.

Similarly, the MR analysis of RLS shows that sequence variants associated with insomnia, neuroticism, smoking behavior, and educational attainment exert causal effect on RLS (β = 1.45, $P = 6.8\times10^{-29}$, β = 0.54, $P = 3.7\times10^{-5}$, β = 0.32, $P = 1.3\times10^{-6}$, β = -0.32, $P = 1.9\times10^{-14}$, respectively). In a recent study, RLS shows strong genetic correlation with neuroticism, smoking behavior, and educational attainment (Didriksen et al., 2020). The MR analysis of TS/Tics using GWAS significant markers of 18 studies highlighted that only the genetic variants associated with insomnia, and educational attainment show causal association with TS/Tics (β = 0.95, $P = 1.3\times10^{-11}$, β = -0.56, $P = 1.0\times10^{-15}$, respectively). To the best of our knowledge, this is the first study evaluating genetic correlation and causal effects of insomnia, and educational attainment PRSs on TS/Tics.

4.7 Understanding causal effect of intracranial volume on ADHD, and Parkinson's disease through GWAS meta-analysis study (paper V)

Paper V reports the largest, to date, GWAS meta-analysis of intracranial volume (ICV) (*N* = 79,174) using summary data from Iceland, UK Biobank, and ENIGMA+EGGC. This study highlights the genetic associations of ICV that impact neurological disorders.

ICV is a quantitative trait that can be measured through magnetic resonance imaging (MRI) / computed tomography or using a tape measure to measure the head circumference (HC) as a proxy measure for ICV. Brain developmental in childhood overlaps with the age at onset of commonly studied neurodevelopmental disorders (ADHD, ASD, Tourette, and OCD) (Hirschtritt et al., 2015). Genetically and phenotypically, ADHD is negatively correlated with ICV (*rg* = -0.23, r = -0.18)(Klein et al., 2019) while Parkinson's disease positively correlates with ICV (*rg* = 0.35, r = 0.08)(Nalls et al., 2019),(Laansma et al., 2021). A key question remains whether sequence variants associated with structural changes in the brain cause neurological disorders, consequent to those structural changes, or alternatively whether genetic predisposition to certain neurological or neurodevelopmental disorders impacts brain structure or development. To understand the impact of ICV on neurodevelopmental disorders we used genetics as a tool to identify the underlying biological relationship. For this, we studied the relationship between ICV and brain function by finding sequence variants impacting brain growth that show causal relationship with neurodevelopmental disorders (**Figure 22**).

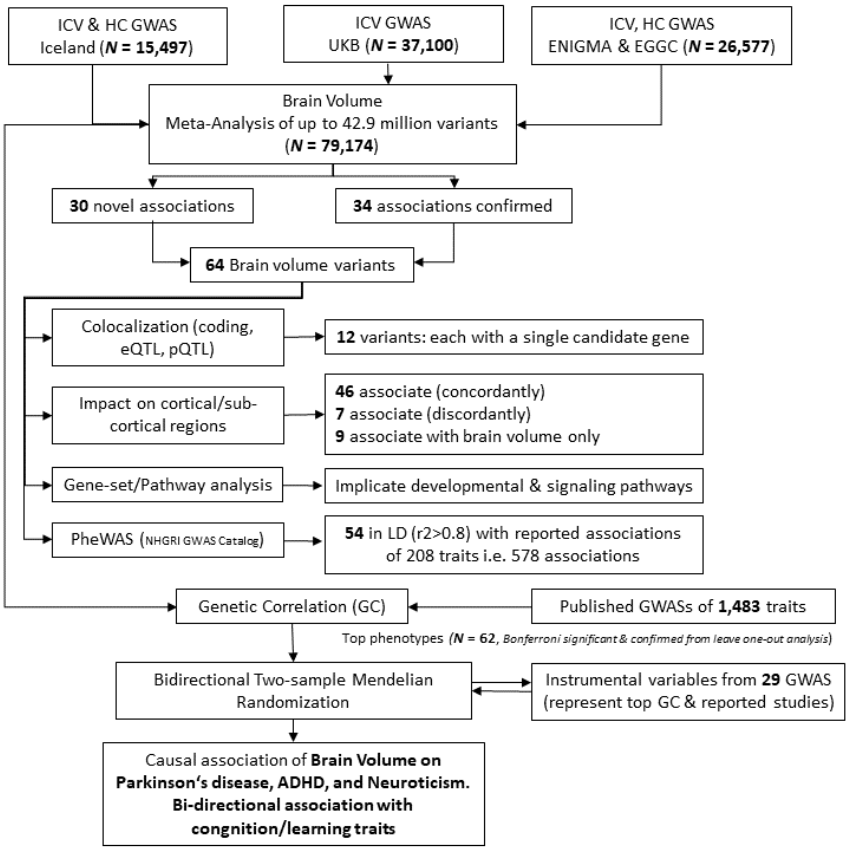


Figure 22: Workflow of the study. A GWAS meta-analysis of ICV by combining GWAS summary data from Iceland, UKB, and ENIGMA+EGGC (total *N* = 79,174) was performed. This study highlights 64 sequence variants that associate with ICV. Based on coding, cis-eQTL, and pQTL analysis 12 genes were identified that exert their impact on ICV. Genetic correlation, gene set enrichment, and phenome scan analyses were performed to identify traits and pathways that correlated with ICV. To understand the underlying biological causal relationship a two-sample bidirectional Mendelian randomization analysis was performed.

4.7.1 Novel variants associated with ICV

The results from three GWAS on ICV (ICV or ICV+HC) were combined as; ICV and HC GWAS from Iceland (N_{ICV+HC} = 15,497), ICV from the UK Biobank (N_{ICV} = 37,100), and ICV and HC from ENIGMA+EGGC (N_{ICV+HC} = 26,577). Altogether, 42.91 million sequence variants were tested for their association with ICV. Therein, a fixed effect meta-analysis was performed by allowing different allele frequencies in each population but assumed to have same effect in each

population. The meta-anlysis highlighted 64 sequence variants, of those 30 are novel, which associate with ICV (**Figure 23**).

The largest effect size on ICV (rs180819997-A, β = -0.191 s.d., P = 2.2×10^{-11}, EAF = 1.05%) is conferred by a novel low frequency variant. Although, this variant also associates with height (β = -0.07, P = 3.31×10^{-9}, N = 511,260) but its effect size (in s.d.) on ICV (adjusted for height) is stronger than that on height ($P_{heterogeneity}$ = 9.01×10^{-5}).

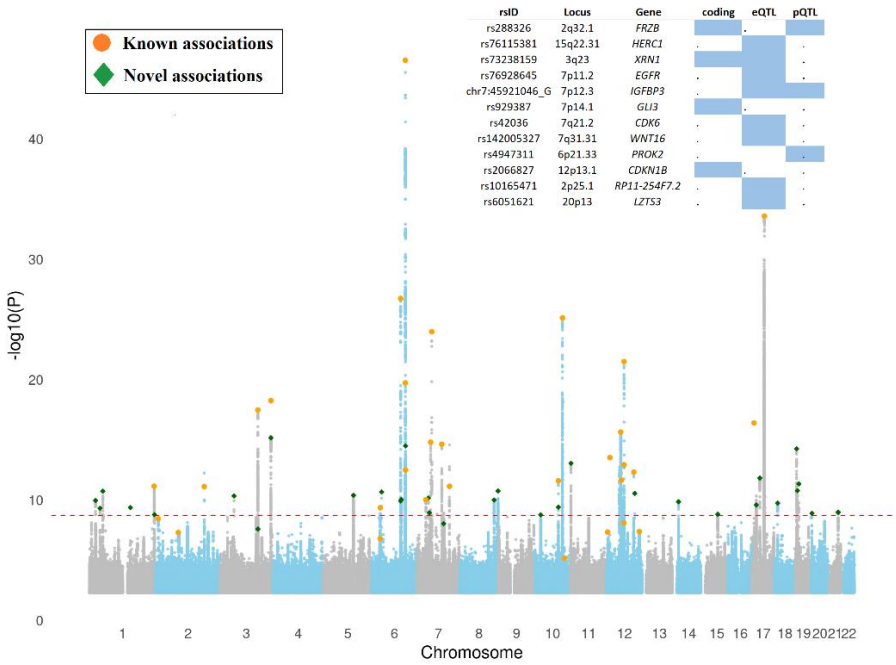


Figure 23: Manhattan-plot showing association results for ICV (*N* = 79,174) with 42.91 million sequence variants (SNPs, In-dels and SVs). Each dot represents a marker tested for association. The *x*-axis shows the chromosomal position of the tested marker, and the *y*-axis the significance (−log$_{10}$*P*) of the observed association. The red dotted line represents suggestive associations set at P = 1.2×10^{-9} (0.05/42.9×10^{6}). Novel associations are highlighted with green diamonds, whereas orange dots represent associations of variants already reported in the scientific literature.

4.7.2 Identification of candidate genes

The transcriptome, proteome, and coding variants analyses were performed to find likely candidate genes that may exert their impact on ICV. The 64 ICV variants (or markers in strong LD, with r^2 > 0.8) were annotated for change in

amino-acid, transcript abundance of colocalized genes, and impact on protein expression in plasma.

Of the 64 ICV variants, 10 colocalized ($r^2 > 0.8$) with coding variants influencing amino-acid change (missense sequence variants) in 10 genes: *MYCL*, *CDKN1B*, *HOOK2*, *FRZB*, *TGOLN2*, *XRN1*, *TNNC1*, *GLI3*, *ZNF789*, and *LRRC24*.

For transcriptome analyses, the RNA expression data from whole blood samples of 13,173 Icelanders as well as the GTEx v8 data were used. Therein, 3,310 genes present within 1mb of the ICV variants were tested for colocalization analysis in 50 tissues and performed 75,728 independent tests (combination of variant × gene × tissue tested, $P_{threshold} < 0.05/75{,}728 = 6.6\times10^{-7}$). The analyses identified 26 ICV variants that colocalized ($r^2 > 0.8$) with top eQTL of a single or multiple genes i.e., 26 ICV variants colocalized with 71 genes.

The proteomic analysis of 4,907 aptamers targeting 4,719 proteins in 35,559 Icelanders identified five ICV variants associating with protein expression of *FRZB, IGFBP3*, *HS6ST2*, *PROK2*, and *CR2* ($P_{threshold} < 0.05/4{,}719/64 = 1.66\times10^{-7}$).

The integrative analysis of three candidate gene studies highlighted that for 12 ICV variants a single gene is implicated: *CDKN1B*, *GLI3*, *FRZB*, *LZTS3*, *XRN1*, *WNT16*, *HERC1*, *RP11-254F7.2*, *IGFBP3*, *EGFR*, *CDK6*, and *PROK2*.

4.7.3 Impact on cortical and sub-cortical regions

The 64 ICV variants were tested for their association with 115 cortical and subcortical volumes (adjusted for ICV) of 37,100 participants who underwent MRI in UK Biobank study. Of the 64 ICV variants, 53 associate with a sMRI trait ($P_{threshold} < 0.05/64/115 = 6.79\times10^{-6}$). Among these 53, six variants show differential effects on local brain volumes as compared to their effect on total ICV (**Figure 24**).

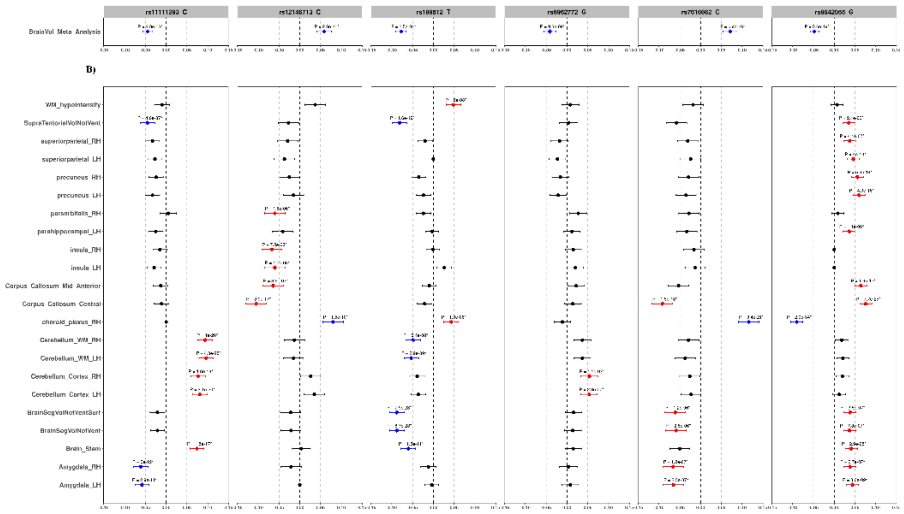


Figure 24: Six ICV variants showing differential effect on local vs. global brain volume. For each variant the effect estimates are plotted for the sMRI traits (on y-axis) that have at-least one divergent effect (red colour highlighted) as compared to their effect on ICV. The concordant associations are highlighted with blue colour while discordant associations are highlighted with red colour (panel 'B'). The p-values are labelled for the significant associations ($P < 0.05/64/115 = 6.79\times10^{-6}$) only. The vertical black dotted line is set at zero representing no effect, while grey dotted lines are shown for better visual comparison of beta estimates at 0.05, and -0.05.

4.7.4 Phenome wide genetic correlation analysis

To find the genetic similarity of ICV with a wide range of phenotypes (disorders, diseases, and traits), the genetic correlation between ICV and 1,483 published GWAS studies was performed using LDSC (B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015). The GC analysis highlights genetic correlation between ICV and 62 of the 1,483 tested phenotypes ($P_{threshold} < 0.05/1,483 = 3.4\times10^{-5}$, **Figure 25**). The positive genetic correlation of ICV with cortical and sub-cortical regions of the brain, Parkinson's disease, educational attainment, and cognitive performance was confirmed (Grasby et al., 2020; Jansen et al., 2020; Nalls et al., 2019) (**Figure 25**). Moreover, the positive genetic correlation was observed between ICV and neonatal traits, social interaction, socioeconomic status, nutritional choice (whole grain), and higher frequency of alcohol intake. Among the negatively correlated traits, the GC between ICV and ADHD and neuroticism were also confirmed (Klein et al., 2019). Additionally, the negative correlation between ICV and having a physical occupation, nutritional choice (white bread), lonliness and sedentary lifestyle was observed.

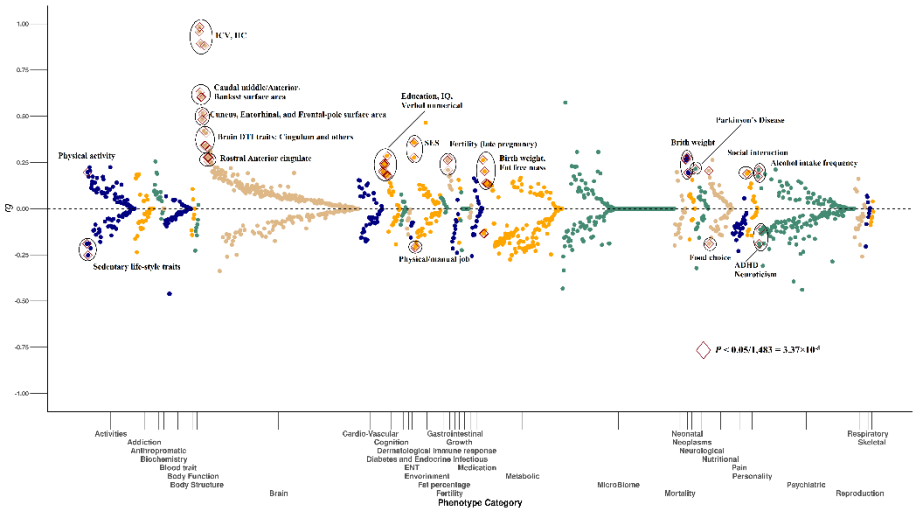


Figure 25: Phenome-wide bivariate genetic correlation between ICV and 1,483 published GWAS studies estimated through LD score regression (B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015). Each dot is an estimate of genetic correlation (*rg*) between ICV GWAS meta-analysis and one of the tested GWAS traits (binned into phenotype categories), where the x-axis represents phenotype (category) and the y-axis shows its genetic correlation (*rg*). The significant associations ($P_{threshold} < 0.05/1{,}483 = 3.37\times10^{-5}$) are highlighted with red-diamond shape.

4.7.5 Bidirectional Mendelian randomization analysis

The GC analysis highlighted that ICV is correlated with wide range of traits, moreover recent studies have also highlighted that common neurological disorders share heritability (Anttila et al., 2018). The genetic correlation simply explores correlation between the tested traits, but the direction of causal relationship remains elusive. To understand the causal relationship, two sample bidirectional Mendelian randomization (MR) analyses were performed to test for causal effect of ICV on 35 traits (represented by genetically correlated and reported associations). The IVs used for MR analyses are sensitive to sample size and strength of association with the predictor phenotype (Morrison et al., 2020).

The 64 ICV variants (instrumental variables) from this study were used as exposure variables to study their effects on correlated traits as an outcome trait: depression (Howard et al., 2019), educational attainment (J. J. Lee et al., 2018b), Parkinson's disease (Nalls et al., 2019), ADHD (D. Demontis et al., 2019), Alzheimer's disease (I. E. Jansen et al., 2019), schizophrenia (M. Lam et al., 2019), bipolar disorder (E. A. Stahl et al., 2019), anorexia (Watson et al., 2019),

Autism spectrum disorder (Grove et al., 2019), OCD (International Obsessive Compulsive Disorder Foundation Genetics & Studies, 2018), Tourette syndrome (D. Yu et al., 2019), migraine (Gormley et al., 2016), epilepsy (International League Against Epilepsy Consortium on Complex, 2018), smoking behaviour (M. Liu et al., 2019), alcohol consumption (M. Liu et al., 2019), big five personality traits (M. Nagel et al., 2018), birth weight (Warrington et al., 2019), and a number of cognitive traits (J. J. Lee et al., 2018a) ($P_{threshold} < 0.05/35 = 1.43\times10^{-3}$**, Table 8**).

The MR analyses show positive causal effects of ICV variants on Parkinson's disease (β = 0.52, $P = 1.22\times10^{-5}$), cognitive traits; verbal numerical reasoning/fluid intelligence (β = 0.139, $P = 3.07\times10^{-10}$), *g* factor (β = 0.102, $P = 6.62\times10^{-5}$), trail making test B (β = -0.093, $P = 3.79\times10^{-5}$), educational attainment (β = 0.073, $P = 9.18\times10^{-8}$), and pairs matching (β = -0.055, $P = 2.11\times10^{-6}$) (**Figure 26, Table 8);** the trail making test B and pairs matching test are scored by how long it takes to complete the task and so a negative value indicates shorter time. The ICV variants show negative causal effects on ADHD (β = -0.203, $P = 6.16\times10^{-4}$), and on neuroticism (β = -0.064, $P = 3.37\times10^{-4}$) (**Figure 26)**. Egger analysis of ICV variants with eight significant traits from inverse variance weighted (IVW) analysis revealed no evidence of variant pleiotropy *i.e.,* the intercepts were not significantly different from zero (**Figure 27)**.

Table 8: Summary of Mendelian randomization analysis using ICV variants as an exposure to test for their causal effect on number of correlated or common brain disorders. outcome.pheno is the trait name on which the effect of ICV was tested, PMID is the reference (PubMed ID or GWAS name) of the outcome trait used for MR analyses, N_{IV} is the count of number of instrumental variables used for analysis (it may vary between study due to availability of the GWAS summary data for respective study), Beta is the causal effect estimated for exposure on outcome, S.E. is the standard error of the causal effect estimate, *P* is the p-value (based on t-distribution) for estimate of causal effect, IVW is inverse variance method used for MR analysis, Egger slope is the estimated causal effect through Egger analyses when intercept is allowed to float, Egger intercept is the estimated intercept based on Egger analyses (a significant non-zero intercept highlights horizontal pleiotropy). **represents un-published data from the GWAS analysis of Icelandic population.

MR Analayis			IVW			Egger Slope			Egger intercept		
outcome.pheno	PMID	N_{IV}	Beta	S.E.	*P*	Beta	S.E.	*P*	Intercept	S.E.	*P*
ADHD	**PMID_30478444**	**60**	**-0.203285**	**0.056176**	**0.000616**	**-0.37691**	**0.19186**	**0.0543**	0.008126	0.009272	0.384
Alzheimer	PMID_30617256	60	0.019914	0.010018	0.0515	0.060004	0.034119	0.0839	-0.00207	0.001698	0.228
Dyscalculia (AMHQ)	Iceland**	57	-0.047004	0.041174	0.258	0.107939	0.127951	0.403	-0.0083	0.006552	0.211
Dyslexcia (ARHQ)	Iceland**	58	-0.071109	0.04484	0.118	-0.03634	0.120962	0.765	-0.00258	0.005536	0.643
Anorexia	PMID_31308545	59	-0.03102	0.07654	0.687	-0.22257	0.273127	0.419	0.009519	0.00983	0.337
Autism	PMID_30804558	59	-0.047966	0.055299	0.389	-0.08543	0.153185	0.579	0.002682	0.009325	0.775
Bipolar disorder	PMID_34002096	49	0.02142	0.056248	0.705	-0.2349	0.181016	0.201	0.013527	0.008613	0.123
Epilepsy	PMID_30531953	65	0.03579	0.031731	0.264	0.075472	0.120473	0.533	9.00E-06	0.004611	0.998
Fluid Intelligence	**UKB_GWAS**	**57**	**0.1421**	**0.018973**	**5.37E-10**	**0.163573**	**0.054665**	**0.00414**	0.00403	0.003287	0.225
Numeric Memory	**UKB_GWAS**	**57**	**0.060903**	**0.019307**	**0.00259**	**0.054859**	**0.061781**	**0.378**	0.002098	0.003157	0.509
Pairs Matching	**UKB_GWAS**	**56**	**-0.05547**	**0.010136**	**1.12E-06**	**-0.05172**	**0.028193**	**0.0721**	-0.0024	0.001713	0.167
Reaction Time	UKB_GWAS	52	-0.036762	0.016877	0.034	-0.07639	0.050356	0.136	0.001857	0.002253	0.414
Symbol Digit	UKB_GWAS	57	0.064248	0.018788	0.00118	0.089478	0.051934	0.0905	0.002842	0.002975	0.344
TMT A	UKB_GWAS	57	-0.056498	0.018212	0.00301	-0.08431	0.06383	0.192	0.001333	0.003051	0.664
TMT B	**UKB_GWAS**	**56**	**-0.091803**	**0.02027**	**3.23E-05**	**-0.09255**	**0.068683**	**0.183**	-0.00056	0.003326	0.866
***g* factor**	**UKB_GWAS**	**58**	**0.108897**	**0.023512**	**2.15E-05**	**0.186343**	**0.065806**	**0.00642**	-9.60E-05	0.003819	0.98
Educational attainment	**PMID_30038396**	**51**	**0.080359**	**0.012112**	**2.24E-08**	**0.056646**	**0.043621**	**0.2**	0.004105	0.002191	0.0669
Insomnia	PMID_30804565	50	-0.04791	0.025494	6.61E-02	-0.04584	0.082744	0.582	-0.00025	0.004104	0.952
Migraine	PMID_27322543	59	0.028628	0.051126	0.578	-0.1294	0.164779	0.436	0.008185	0.00751	0.28
Multiple sclerosis	not_published*	65	0.072667	0.093613	0.44	-0.3156	0.525301	0.55	0.025699	0.025235	0.312
Agreeableness	PMID_21173776	59	0.0766	0.02405	0.00233	0.032154	0.06851	0.641	0.005528	0.003799	0.151
Conscientiousness	GCST006326	63	0.029448	0.025301	0.249	-0.08078	0.083015	0.334	0.005759	0.004015	0.157
Extraversion	PMID_26362575	63	-0.009835	0.016211	0.546	0.044473	0.056609	0.435	-0.00358	0.002675	0.185
Openess	GCST000922	63	0.075045	0.031696	0.021	0.126615	0.10597	0.237	-0.00302	0.005335	0.574
Neuroticism	**PMID_29942085**	**51**	**-0.064406**	**0.016732**	**0.000337**	**-0.15044**	**0.053732**	**0.0073**	0.004327	0.0028	0.129
OCD	PMID_28761083	59	-0.001568	0.002214	0.482	-0.00718	0.0066	0.281	0.000573	0.000614	0.355
Parkinson	**PMID_31701892**	**53**	**0.537002**	**0.105162**	**4.74E-06**	**0.844878**	**0.341779**	**0.0168**	-0.01486	0.01778	0.407
Depression	PMID_30718901	59	-0.060951	0.024797	0.017	-0.06639	0.083606	0.43	-0.00097	0.004017	0.81
Schizophrenia	PMID_25056061	51	-0.065141	0.069861	0.356	0.10959	0.266168	0.682	-0.00899	0.013134	0.497
Tourette	PMID_30818990	56	-0.076061	0.104237	0.469	-0.24776	0.357358	0.491	0.009115	0.017334	0.601
Smoker current Vs Former	**PMID_30643251**	**53**	**-0.062043**	**0.019036**	**0.00197**	**-0.08033**	**0.062126**	**0.202**	0.000826	0.003125	0.793
Smoker Ever vs Never	PMID_30643251	53	-0.029511	0.024413	0.232	0.065932	0.083892	0.436	-0.00496	0.00421	0.244
Drinks Per Week	PMID_30643251	53	0.022677	0.011469	0.0533	0.013592	0.039924	0.735	0.000764	0.002005	0.705
Birth Weight maternal	PMID_31043758	60	0.123313	0.041854	0.0048	0.097262	0.133326	0.469	0.001856	0.005911	0.755
Birth Weight child	PMID_31043758	60	0.12328	0.049007	0.0146	0.070863	0.175127	0.687	0.003784	0.007679	0.624

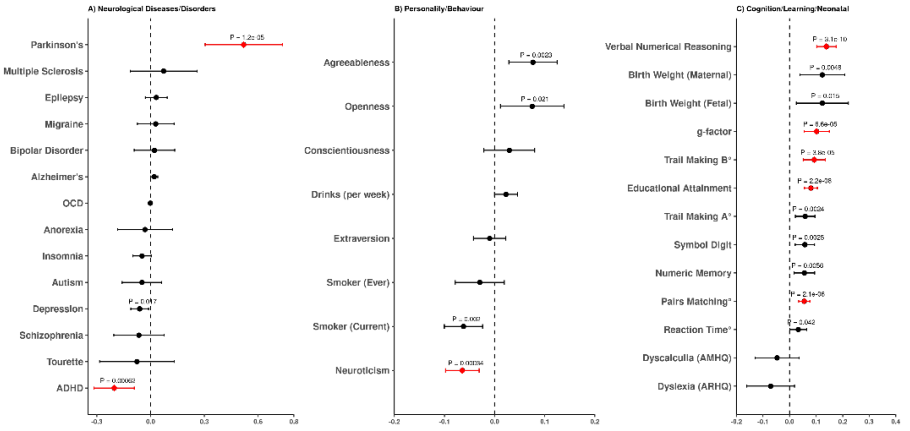


Figure 26: Causal association of instrumental variants from ICV on 34 tested traits (A) neurological diseases/disorders (B) personality/behavioural traits (**C**) cognitive/learning/birth weight traits**.** The analysis was performed using a two-sample Mendelian randomization (MR) approach, the instrumental variables and their effect sizes are based on results for ICV variants compared to their effects from the largest available studies of the genetically correlated traits. IVW (inverse variance weighted) method was used to estimate the causal effect, additionally Egger analysis was performed to detect whether IVW estimates are biased i.e., intercept is different from zero. The Bonferroni significant associations ($P < 0.05/34 = 1.47\times10^{-3}$) are highlighted with red-color, '*or*' refers to traits for which effect estimates were flipped for better representation.

For the reverse causal analysis, we tested GWAS significant variants of 29 studies as exposures (IVs) to explore the potential causal effects on ICV from this study ($P_{threshold} < 0.05/29 = 1.72\times10^{-3}$, **Table 9**). The exposures of Parkinson's disease, neuroticism, and migraine on ICV had a nominally significant effect in our MR analysis (not significant after adjusting for multiple testing and in leave-one-sample-out analysis). The exposures of birth weight, insomnia, cognitive and learning traits show causal effect on ICV ($\beta_{birth\text{-}weight} = 0.216$, $P_{birth\text{-}weight} = 1.16\times10^{-6}$; $\beta_{insomnia} = -0.192$, $P_{insomnia} = 7.07\times10^{-6}$; $\beta_{education} = 0.264$, $P_{education} = 4.11\times10^{-33}$; $\beta_{cognitive\ performance} = 0.190$, $P_{cognitive\ performance} = 8.62\times10^{-9}$). Thus, only cognition and learning traits show bi-directional causal relationship with ICV.

Table 9: Summary of Mendelian randomization analysis using instrumental variables of correlated studies as an exposure to test for their causal effects on ICV. Exposure.pheno refers to effect estimates from correlated study as exposure trait, PMID is the reference (pubmed ID or GWAS name) of the exposure trait used for MR analyses, N_{IV} is the number of instrumental variables used for analysis (number of independently associated GWAS significant variants), Beta is the causal effect estimated for exposure on outcome, S.E. is the standard error of the causal effect estimate, *P* is the p-value (based on t-distribution) for estimate of causal effect, IVW is inverse variance method used for MR analysis, Egger slope is the estimated causal effect through Egger analyses when intercept is allowed to float, Egger intercept is the estimated intercept based on Egger analyses (a significant non-zero interceprt highlights horizontal pleiotropy).

MR Analayis			IVW			Egger Slope			Egger intercept		
exposure.pheno	PMID_(exposure)	N_{IV}	Beta	S.E.	*P*	Beta	S.E.	*P*	Intercept	S.E.	*P*
ADHD	PMID_30478444	9	0.01907	0.024164	0.456	0.048963	0.103531	0.653	-0.002758	0.009683	0.785
Age started smoking	PMID_30643251	10	-0.12611	0.104113	0.257	-0.124167	0.583841	0.832	-0.000132	0.010534	0.99
Alzhemier	PMID_30617256	32	-0.040861	0.05912	0.495	-0.020789	0.08779	0.814	-0.000682	0.002243	0.763
Anorexia Nervosa	PMID_31308545	8	-0.017617	0.028183	0.552	0.006332	0.119691	0.96	-0.002191	0.011096	0.85
Bipolar disorder	PMID_31043756	139	0.004025	0.009726	0.68	-0.007707	0.037015	0.835	0.000913	0.002578	0.724
Cigerates per day	PMID_30643251	55	-0.051772	0.046163	0.267	-0.100341	0.096972	0.305	0.001878	0.002528	0.461
Cognitive Performance MTAG	**PMID_30038396**	**653**	**0.228407**	**0.020384**	**9.13e-27**	**0.318795**	**0.092807**	**0.00063**	**-0.000759**	**0.001642**	**0.644**
Cognitive Performance	**PMID_30038396**	**225**	**0.190219**	**0.031779**	**8.62e-09**	**0.300772**	**0.157907**	**0.0581**	**-0.001246**	**0.003304**	**0.706**
Depression	PMID_30718901	97	-0.011862	0.03741	0.752	0.281007	0.17043	0.102	-0.006431	0.003644	0.0808
Drinks per week	PMID_30643251	97	-0.005365	0.075899	0.944	0.045456	0.167369	0.787	-0.001313	0.002194	0.551
Education Years COJO	**PMID_30038396**	**447**	**0.284993**	**0.034135**	**8.88e-16**	**0.236637**	**0.132796**	**0.0754**	**0.000893**	**0.001531**	**0.56**
Education years MTAG	**PMID_30038396**	**1599**	**0.290064**	**0.019958**	**4.84e-45**	**0.419695**	**0.078008**	**8.55e-08**	**-0.000924**	**0.000836**	**0.269**
Education Years	**PMID_30038396**	**1252**	**0.264325**	**0.021415**	**4.11e-33**	**0.422877**	**0.083669**	**4.97e-07**	**-0.001353**	**0.000949**	**0.154**
Epilepsy	PMID_30531953	10	-0.00253	0.001534	0.134	-0.051247	0.048576	0.322	0.085364	0.086167	0.351
Highest Math Ability	**PMID_30038396**	**361**	**0.231528**	**0.03009**	**1.38e-13**	**0.400417**	**0.16249**	**0.0142**	**-0.001933**	**0.00274**	**0.481**
Highest MATH MTAG	**PMID_30038396**	**1295**	**0.253127**	**0.018455**	**4.46e-40**	**0.42859**	**0.074006**	**8.76e-09**	**-0.001874**	**0.000979**	**0.0558**
Intelligence	**PMID_29942086**	**241**	**0.195824**	**0.030587**	**8.19e-10**	**0.331664**	**0.154274**	**0.0326**	**-0.001666**	**0.003121**	**0.594**
Insomnia	**Kyoko_Watanabe_medRxiv_2020**	**780**	**-0.191448**	**0.042341**	**7.07e-06**	**-0.036237**	**0.167758**	**0.829**	**-0.001461**	**0.001208**	**0.227**
MATH ability MTAG	**PMID_30038396**	**860**	**0.212092**	**0.021926**	**4.45e-21**	**0.452228**	**0.095963**	**2.85e-06**	**-0.002868**	**0.001344**	**0.0331**
Math Ability	**PMID_30038396**	**611**	**0.168358**	**0.024249**	**9.85e-12**	**0.331777**	**0.110681**	**0.00283**	**-0.001902**	**0.001651**	**0.25**
Migraine	Heidi_etal_medRxiv_2021	118	-0.042404	0.017356	0.016	0.042244	0.049771	0.398	-0.004544	0.002284	0.049
Multiple sclerosis	PMID_31604244	275	-0.005685	0.006091	0.351	0.029302	0.019694	0.138	-0.003272	0.00167	0.051
Neuroticism	PMID_29942085	135	-0.126352	0.049029	0.011	-0.648068	0.262893	0.015	0.007919	0.004009	0.0503
Birth Weight Maternal	**PMID_31043758**	**72**	**0.158687**	**0.044611**	**0.000673**	**-0.093567**	**0.163118**	**0.568**	**0.007824**	**0.004375**	**0.078**
Birth Weight	**PMID_31043758**	**144**	**0.216029**	**0.042504**	**1.16e-06**	**0.545375**	**0.159052**	**0.000796**	**-0.007825**	**0.003745**	**0.0385**
Parkinson's disease	PMID_31701892	90	0.039759	0.011887	0.00121	0.002327	0.02603	0.929	0.004999	0.002609	0.0586
Schizophrenia	PMID_25056061	111	0.00678	0.011236	0.548	0.056883	0.049883	0.257	-0.003895	0.00383	0.311
Smoker current vs former	PMID_30643251	24	0.102715	0.054126	0.0709	0.08245	0.15068	0.59	0.001167	0.005786	0.842
Smoker Ever vs never	PMID_30643251	369	0.043947	0.024372	0.0722	0.096762	0.113857	0.396	-0.00105	0.002184	0.631

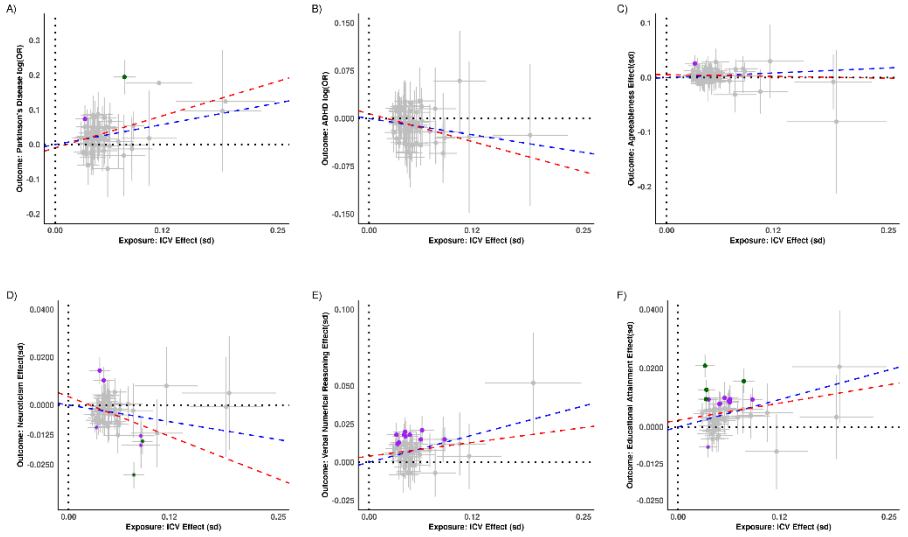


Figure 27: Effect vs effect plots of top associations from MR analysis. On x-axis are effect size for ICV and on y-axis (not always symmetric around '0') the effect size or; (**A**) Parkinson's disease as log (odds ratio) (**B**) ADHD as log (odds ratio) (**C**) Agreeableness as beta in S.D. (**D**) Neuroticism as beta in S.D., (**E**) Verbal numerical reasoning as beta in S.D., and (**F**) Educational attainment as beta in S.D. All effects are plotted for alleles with increasing ICV. Blue line represents the estimated slope from IVW (inverse variance weighted regression), and red line is estimated from MR Egger analysis including the intercept. Green dots represent conventional GWAS associations ($P < 5.0\times10^{-8}$) for respective y-axis trait, while purple dots are Bonferroni significant associations ($P < 0.05/64 = 7.8\times10^{-4}$) for respective the y-axis trait.

4.7.6 Conclusion

Total ICV can capture structural variations in the brain. The discovery of sequence variants associated with brain structures may help to understand its functioning. So far, limited number of sequence variants associating with ICV have been identified. In this analysis (using ICV of 79,174 participants), 64 associations were highlighted, including 30 novel variants. The largest impact on ICV is exerted by a low frequency variant, located at 6p21.2, (rs180819997-A, β = -0.191 s.d., $P = 2.2\times10^{-11}$). Polygenic risk score of ICV (p-value threshold = 0.1) explains up to 8.78% of phenotypic variance while 64 ICV variants explain 5.0% variance in ICV.

Biological annotation of the 64 ICV variants through transcriptome, proteome, and coding variant analyses highlighted that 12 of 64 ICV variants exert their impact through a single candidate gene (including *GLI3*, *CDK6*, and *FRZB)*. *GLI3* regulates early developmental mechanisms and rare mutations in *GLI3* associate with premature fusing of the skull (Hurst et al., 2011). In this study, a common sequence variant in *GLI3* (p.Asp1137Asn) associates with larger ICV and may be involved in delayed fusing of the skull sutures. Recessive mutations in *CDK6* associate with microcephly (Hussain et al., 2013), a common sequence variant identified in this study suggests that lowered *CDK6* expression is associated with smaller ICV. Moreover, *FRZB* has been shown to play an important role in osteogenesis (Jin et al., 2016; Loughlin et al., 2004), which is in line with the common coding variant, p.His488Gln, which associates with a larger ICV and higher FRZB protein expression. These findings highlight possible role of *GLI3*, *CDK6*, and *FRZB* during skull/brain development.

Many ICV variants also associate with personality/cognitive/learning traits, cardiovascular disorders, neurological and autoimmune disorders. These associations indicate a shared etiology between ICV and a number of phenotypes (diseases, disorders, and or traits). More specifically, six of the 64 ICV variant show divergent effect on volumes of some of the cortical and subcortical regions. The divergent associations may help to understand brain region specific roles involving horizontal pleiotropy. Among those, a common sequence variant, the 17q21.3 inversion, is notable. The 17q21.31 inversion, first identified in the Icelandic population(Stefansson et al., 2005), is under positive selection and is known to associate with many neurological disorders, personality traits, and ICV. The inversion polymorphism located at 17q21.31 has two haplotypes in Caucasian populations, H1 and H2. H1 associates with Parkinson's disease (Nalls et al., 2019) and a larger ICV. One of the genes affected by the inversion polymorphism is *MAPT*, which is a candidate gene in Parkinson's disease, based on its involvement with tauopathies. H2, the inverted haplotype, associates with smaller ICV, neuroticism (M. Nagel et al., 2018) and negatively associates with cognitive traits (J. J. Lee et al., 2018a).

The GC of ICV compared with 1,483 published GWAS studies confirmed many known correlations, such as with Parkinson's disease, ADHD, cognitive/learning traits, educational attainment, neuroticism, and cortical and sub-cortical regions. Additionally, the genetic correlation between ICV and neonatal traits, socio-economic status, environmental traits, sedentary lifestyle, having a physical occupation, and higher frequency of alcohol intake was observed. The genetic

correlation can identify the strength of correlation (positive or negative) between the tested traits, but the underlying true causal relationship remains elusive.

To understand the causal relationship between ICV and traits or disorders genetically correlated to or sharing heritability with ICV (Anttila et al., 2018), two-sample bidirectional Mendelian randomization (MR) analyses were performed. For this, robustly associated GWAS significant sequence variants were used as instrumental variables (IVs), for forward and reverse MR analyses. The causal analyses of 64 ICV variants on 35 phenotypes (diseases, disorders, or traits) highlighted that ICV has positive causal effect on Parkinson's disease ($\beta = 0.52$, $P = 1.22\times10^{-5}$), and negative causal effect on ADHD ($\beta = -0.203$, $P = 6.16\times10^{-4}$), and neuroticism ($\beta = -0.064$, $P = 3.37\times10^{-4}$). The reverse MR analyses did not reveal significant evidence of Parkinson's disease, ADHD, or neuroticism on ICV.

Most noticeable is the negative causal effect of insomnia on ICV ($\beta_{insomnia} = -0.192$, $P_{insomnia} = 7.07\times10^{-6}$). Insomnia has a strong genetic component where 780 sequence variants have been reported to associate with insomnia. Moreover, insomnia also shows positive causal effect on ADHD (P. R. Jansen et al., 2019). In this study, ICV shows negative causal effect on ADHD and insomnia has negative causal effect on ICV. These findings strongly implicate that genetic predisposition to insomnia negatively impacts brain development which in turn affects ADHD.

The bidirectional MR analyses between ICV and cognitive/learning traits is inconclusive, as strong bidirectional relationship was observed between these traits. Further studies are required to dissect the causal pathways of ICV with cognitive and learning traits.

This study used largest GWAS meta-analysis of ICV to highlight 64 associations and implicated 12 genes which likely impact ICV. The causal analysis uncovered underlying biological relationship between ICV and a neurodevelopmental (ADHD) disorder, and a neurodegenerative disease (Parkinson's). Our findings highlight that either changes in ICV show causal effect on ADHD and Parkinson's disease or ICV closely correlates with a confounder that may explain this causal relationship. Importantly, this study helped to understand that genetic predisposition to insomnia exerts its impact on ADHD through a causal affect on ICV. Furthermore, based on ICV and genotype data of 79,174 participants, 5.0% of the phenotypic variance is attributed to 64 sequence variants. This study further helps to understand the relationship between brain structure and brain function.

5 Discussion

The aim of the genetic studies described in this thesis was to dissect the genetic architecture of human intracranial volume and to understand the genetic aetiology of impulsivity-compulsivity disorders (TS, Tics, OCD, ADHD, and RLS). So far only a few sequence variants are unequivocally associated with these disorders (D. Demontis et al., 2019; Didriksen et al., 2020; Simon Haworth et al., 2019; Jansen et al., 2020; Smit et al., 2020; D. Yu et al., 2019). This may partially be attributed to heterogenic, and/or comorbid nature of these disorders or limitations of underpowerd studies. Thus, we searched for association between rare and common variants with ICV and impulsivity-compulsivity disorders.

Candidate CNV meta-analysis confirmed that *AADAC* deletion is a risk factor for TS. Rare, recurrent, so called, neuropsychiatric CNVs have been shown to associate with increased risk for ASD, developmental disorders, and schizophrenia (Ingason et al., 2011; Kirov et al., 2014; Malhotra & Sebat, 2012; Morrow, 2010; Pinto et al., 2014; Stefansson et al., 2008). This thesis shows that neuropsychiatric CNVs confer high risk of ADHD. Furthermore, 17q12 duplication (a neuropsychiatric CNV) is highlighted as a risk factor for TS. These findings highlighted that ADHD and TS share rare genetic risk factors with ASD and schizophrenia. The neuropsychiatric CNVs are known to exhibit dose-dependent effects on human brain structural and functional alteration (Stefansson et al., 2014). It is likely that dose-sensitive genes affected by recurrent CNV loci in combination with polygenic risk scores of respective disorder(s) or brain structure(s) may affect disorders differently. Future studies are required to better understand such complex interplay of rare and common risk variants.

Five GWAS meta-analyses of TS, Tics, OCD, RLS, and ICV were performed and identified strong associations with RLS and ICV. The GWAS meta-analysis of TS (cases = 4,819) didn't find any sequence variant associated with TS (D. Yu et al., 2019). While, the study highlighted polygenic architecture of TS where TS symptoms and severity increased with higher PRS score (D. Yu et al., 2019). To detect a common variant (above 1% frequency) with small effect estimate (OR less the 1.2), a large effective sample size is required (Crouch & Bodmer, 2020). For TS, a larger effective sample size is expected to uncover individual association signals as was the experience of the schizophrenia, and ADHD

GWAS. The GWAS for Tics highlighted a rare frame-shift variant in *EGFL7* that associates with increased risk of Tics disorder. Replication of this signal in an independent population is required to confirm this association. *EGFL7* is involved in angiogenesis and higher expression was observed in mice autoimmune encephalomyelitis (EAE) where it may be used to reduce inflammation (Larochelle et al., 2018). The GWAS meta-analysis of OCD didn't find any significant associations. Like TS, a larger sample is required for a GWAS meta-analysis to uncover novel associations.

The combined GWAS meta-analysis of RLS from six populations (cases = 10,257) highlighted 23 GWAS significant associations. Cis-colocalization analysis of those 23 sequence variants implicated five genes (*RANBP17*, *CASC16*, *HBOX2*, *MAP2K5*, and *SKOR1*) potentially involved in RLS aetiology. These genes provide suggestive markers to investigate a role for druggable targets in RLS treatment. Among these, rs10068599-T associates with increased risk for RLS and lower expression of *RANBP17* in brain subcortical regions, mainly in the basal ganglia. Basal ganglia is involved in modulating emotional and motor output (Pierce & Peron, 2020). Moreover, Parkinson's disease (a movement disorder) is characterized by a loss of dopaminergic innervation in the basal ganglia impacting motor symptoms (Neumann et al., 2018). RLS exhibits involuntary urges to move legs while in Parkinson's disease the voluntary control of movements is compromised. Future studies may help to understand the role of *RANBP17* in RLS aetiology and uncover more signals and to yield deeper insights into the disease biology.

Genome-wide genetic correlation analysis of five impulsivity-compulsivity disorders found 59 common genetically correlated traits shaping shared genetic architecture. Hierarchical clustering of these correlated traits highlighted five latent clusters ranging from psychiatric, emotional, cognitive, lifestyle related traits, and pain disorders. As the genetic correlation may or may not exhibit causation, the Mendelian randomization analyses were performed (using GWAS markers from 18 studies) to test for causal effect of common genetically correlated traits on five impulsivity-compulsivity disorders. Among these, the insomnia associated sequence variants exerted the strongest impact on all tested disorders ($\beta > 0.85$, $P < 6.7\times10^{-10}$). This implies that insomnia has a causal effect on the neurological disorders. Notably, a strong genome-wide genetic correlation between insomnia and neurological disorders was also observed. Future studies are required to understand whether there is a bi-directional effect between neurological disorders and insomnia (sleep disturbances).

The brain is in growth phase during childhood and variation in ICV associates with several neurological disorders. It is likely that sequence variants exert their effect on neurological disorders through their impact on structural variations in ICV. The largest, to date, GWAS meta-analysis of ICV found 64 variants explaining 5% of variance. The genetic correlation of ICV compared against 1,480 GWAS studies, found 62 traits correlated with ICV: including ADHD, Parkinson's disease, cognition and learning traits, neuroticism, and socio-economic status. Parkinson's disease cases have greater ICV whereas ADHD cases have smaller ICV than controls (at phenotypic level). Bidirectional MR analyses of ICV vs correlated studies revealed that ICV either has a causal effect on a neurodevelopmental disorder (ADHD) as well as on a neurodegenerative disease (Parkinson's) or these causal relationships might be driven by traits closely correlated with ICV. These findings underscore the potential of using the simple measure of ICV combined with genetics to further investigate brain structure-function relationships. Future studies may focus on understanding the closely correlated traits between ICV, ADHD, and Parkinson's disease.

6 Conclusions

Human genetic research is aimed at understanding the evolution, diversity, and aetiology of disease and, from a medical perspective, to identify druggable targets which are helpful for better treatment. The neuropsychiatric disorders show high heritability, polygenicity, complex interplay of polygenic markers, and gene-environment interaction. Large studies are required to discover genetic associations and to uncover underlying genetic architecture.

The candidate gene approach aims to test a gene based on the function of its protein or previous association results. In 1st study, a suggestive finding (Sundaram et al., 2010) was followed up in a large sample, from seven European countries, of TS patients and population-based controls. The *AADAC* deletion was identified as a risk factor for TS ($P = 1.7\times10^{-5}$, OR = 1.58). AADAC exhibited higher expression in various brain tissues suspected to be involved in neuronal function. While our results robustly replicate the previous observations, further studies are required to improve the risk estimate but also to investigate AADAC's role in the TS aetiology. Furthermore, through a larger GWAS meta-analysis of TS from Caucasian populations, it was shown that TS is highly polygenic in nature and a TS PRS associates with increased risk of TS and tics symptom severity ($P = 5.3\times10^{-9}$, OR = 1.33). The polygenic nature and identification of developmental circuit pathways suggested that TS is a complex developmental circuit disorder affecting motor, cognitive, and behavioural controls. Larger GWAS studies, functional and structural MRI analyses, and causal analysis, are required to further understand these preliminary findings of TS.

ADHD is a highly comorbid, heterogenous, and disabling disorder. Studies have demonstrated high heritability and polygenicity (mostly tagged by common sequence variants) of ADHD. Rare coding variants exert large impact but identifying them is constrained by inadequate sample sizes. Neuropsychiatric CNVs, a group of 19 rare, recurrent CNVs, have been shown to associate with increased risk of autism, and schizophrenia, and to affect cognition in individuals without neurodevelopmental or psychiatric disorders. ADHD is comorbid with these disorders, but their underlying shared genetics remain elusive. Based on a large meta-analysis of ADHD cases and controls from Iceland and Norway, it was shown that neuropsychiatric CNVs also confer risk of ADHD

(OR = 2.43, $P = 1.6 \times 10^{-21}$). Moreover, the 17q12 duplication, a neuropsychiatric CNV, associates with increased risk of TS ($P = 8.7\times10^{-6}$, OR = 10.43). However, the 17q12 duplication did not associate with ADHD. These findings confirmed pleiotropic effects of neuropsychiatric CNVs and suggests shared aetiology of ADHD, TS, autism, and schizophrenia. Furthermore, the association of rare neuropsychiatric CNVs provides evidence for ADHD, TS, ASD and schizophrenia being related neurodevelopmental disorders rather than distinct entities.

RLS is a complex polygenic sensorimotor disorder strongly influenced by lifestyle. Based on GWAS meta-analysis of RLS from six populations, including more than 10,000 cases and 470,000 controls, 23 independent GWAS significant sequence variants were identified. The cis-eQTL colocalization analysis of RLS variants using eQTL data from 49 tissues, implicated five genes (*RANBP17*, *CASC16*, *HBOX2*, *MAP2K5*, and *SKOR1*) as possibly involved in RLS aetiology. These genes provide suggestive markers to investigate for role of druggable targets in RLS treatment. The polygenic risk score and genetic correlation analyses of RLS confirmed prior epidemiological findings that implicate obesity, smoking and high alcohol consumption as risk factors for RLS. The 23 GWAS signals explain less than 1% of the variance in RLS, which suggests larger GWAS studies are required to uncover more signals and to yield deeper insights into the disease biology.

Common neurodevelopmental disorders (TS, Tics, ADHD, and OCD) and common involuntary movement disorders (RLS) share phenotypic overlap including involuntary motor, and compulsion components. PRS analysis confirmed their polygenic nature while pairwise genetic correlation analyses highlighted genetic overlap. Yet their underlying genetic architecture remains elusive. Based on genetic correlation analyses of these disorders compared against 1,140 published GWAS studies identified 59 common traits. The hierarchical clustering of correlated traits highlighted five latent clusters: (1) neuropsychiatric or neurotic disorders, (2) emotional disorders, (3) peripheral and muscular pain, (4) obesity / poor lifestyle, and, (5) cognition / learning traits. Genetic correlation may or may not indicate causation. To infer causality, GWAS significant variants (as instrumental variables - IVs) from 18 genetic studies were used in Mendelian randomization analysis to test for their causal effect on TS, Tics, ADHD, OCD, and RLS. Among the tested IVs, the insomnia associated sequence variants exerted the strongest impact on all tested traits/disorders ($\beta > 0.85$, $P < 6.7\times10^{-10}$). This implies that insomnia has a causal effect on the neurological disorders. Notably, a strong genome-wide genetic

correlation between insomnia and neurological disorders was also observed. Future studies are required to understand whether there is a bi-directional effect between neurological disorders and insomnia (sleep disturbances).

The structural variations seen in brain may associate with neurological disorders. Moreover, sequence variants may exert their effect on neurodevelopmental disorders through their impact on brain growth. Variation in brain structure can help to investigate brain structure-function relationships. ICV was used as a measure of total brain size. GWAS meta-analysis (from four studies) of ICV found 64 variants explaining 5% of variance. The genetic correlation of ICV compared against 1,480 GWAS studies, found 62 traits correlated with ICV: including ADHD, Parkinson's disease, cognition and learning traits, neuroticism, socio-economic status, birth weight and measures of cortical and subcortical regions. Furthermore, Parkinson's disease cases have greater ICV whereas ADHD cases have smaller ICV than controls (at phenotypic level). Bidirectional MR analyses of ICV compared with 34 GWAS studies, revealed that ICV either has a causal effect on a neurodevelopmental disorder (ADHD) as well as on a neurodegenerative disease (Parkinson's) or these causal relationships might be driven by traits closely correlated with ICV. These findings underscore the potential of using the simple measure of ICV combined with genetics to further investigate brain structure-function relationships.

References

Affymetrix, I. (2013). Axiom® genotyping solution data analysis guide. In: URL.

Affymetrix, I. (2015). UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory. *URL*.

Affymetrix, I. (2017). UKB_WCSGAX: UK Biobank 500K Samples

Processing by the Affymetrix1 Research Services

Laboratory. *URL*.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664. doi:10.1101/gr.094052.109

Allen, R. P., Picchietti, D. L., Garcia-Borreguero, D., Ondo, W. G., Walters, A. S., Winkelman, J. W., . . . Lee, H. B. (2014). Restless legs syndrome/Willis-Ekbom disease diagnostic criteria: updated International Restless Legs Syndrome Study Group (IRLSSG) consensus criteria–history, rationale, description, and significance. *Sleep Med, 15*(8), 860-873. doi:10.1016/j.sleep.2014.03.025

Biobank., U. https://www.ukbiobank.ac.uk/scientists-3/genetic-data/.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . O'Connell, J. J. B. (2017). Genome-wide genetic data on~ 500,000 UK Biobank participants. 166298.

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol, 11*(4), e1004219.

Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *Am J Hum Genet, 93*(4), 687-696. doi:10.1016/j.ajhg.2013.09.002

Di Angelantonio, E., Thompson, S. G., Kaptoge, S., Moore, C., Walker, M., Armitage, J., . . . Danesh, J. (2017). Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet, 390*(10110), 2360-2371. doi:10.1016/s0140-6736(17)31928-1

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics, 30*(9), 1266-1272. doi:10.1093/bioinformatics/btu014

Endres-Dighe, S. M., Guo, Y., Kanias, T., Lanteri, M., Stone, M., Spencer, B., . . . Busch, M. P. (2018). Blood, sweat, and tears: Red Blood Cell-Omics study objectives, design, and recruitment activities. *Transfusion*. doi:10.1111/trf.14971

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Hussain, M. S., Baig, S. M., Neumann, S., Peche, V. S., Szczepanski, S., Nurnberg, G., . . . Noegel, A. A. (2013). CDK6 associates with the centrosome during mitosis and is mutated in a large Pakistani family with primary microcephaly. *Hum Mol Genet, 22*(25), 5199-5214. doi:10.1093/hmg/ddt374

Ji, S. G., Juran, B. D., Mucha, S., Folseraas, T., Jostins, L., Melum, E., . . . Anderson, C. A. (2017). Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet, 49*(2), 269-273. doi:10.1038/ng.3745

Jin, C., Jia, L., Huang, Y., Zheng, Y., Du, N., Liu, Y., & Zhou, Y. (2016). Inhibition of lncRNA MIR31HG Promotes Osteogenic Differentiation of Human Adipose-Derived Stem Cells. *Stem Cells, 34*(11), 2707-2720. doi:10.1002/stem.2439

Jonsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., . . . Stefansson, K. (2017). Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci Data, 4*, 170115. doi:10.1038/sdata.2017.115

Kanias, T., Lanteri, M. C., Page, G. P., Guo, Y., Endres, S. M., Stone, M., . . . Gladwin, M. T. (2017). Ethnicity, sex, and age are determinants of red blood cell storage and stress hemolysis: results of the REDS-III RBC-Omics study. *Blood Adv, 1*(15), 1132-1141. doi:10.1182/bloodadvances.2017004820

Larochelle, C., Uphaus, T., Broux, B., Gowing, E., Paterka, M., Michel, L., . . . Prat, A. (2018). EGFL7 reduces CNS inflammation in mouse. *Nature communications, 9*(1), 1-12.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Y, A. R., H, K. F., . . . A, L. P. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet, 48*(11), 1443-1448. doi:10.1038/ng.3679

Loughlin, J., Dowling, B., Chapman, K., Marcelline, L., Mustafa, Z., Southam, L., . . . Corr, M. (2004). Functional variants within the secreted frizzled-related protein 3 gene are associated with hip osteoarthritis in females. *Proc Natl Acad Sci U S A, 101*(26), 9757-9762. doi:10.1073/pnas.0403456101

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . Haplotype Reference, C. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet, 48*(10), 1279-1283. doi:10.1038/ng.3643

Olafsdottir, T. A., Theodors, F., Bjarnadottir, K., Bjornsdottir, U. S., Agustsdottir, A. B., Stefansson, O. A., . . . Eyjolfsson, G. I. (2020). Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. *Nature Communications, 11*(1), 1-11.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet, 38*(8), 904-909. doi:10.1038/ng1847

Spencer, B. R., Kleinman, S., Wright, D. J., Glynn, S. A., Rye, D. B., Kiss, J. E., . . . Cable, R. G. (2013). Restless legs syndrome, pica, and iron status in blood donors. *Transfusion, 53*(8), 1645-1652. doi:10.1111/trf.12260

Timmer, T. C., de Groot, R., Habets, K., Merz, E. M., Prinsze, F. J., Atsma, F., . . . van den Hurk, K. (2019). Donor InSight: characteristics and representativeness of a Dutch cohort study on blood and plasma donors. *Vox Sang, 114*(2), 117-128. doi:10.1111/vox.12731

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., . . . Do, R. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics, 97*(4), 576-592.

Yuelong Guo, M. B., Mark Seielstad, Stacy Endres-Dighe, Connie M. Westhoff, Brendan Keating, Carolyn Hoppe, Aarash Bordbar, Brian Custer, Adam S. Butterworth, Tamir Kanias, Alan E. Mast, Steve Kleinman, Yontao Lu, and Grier P. Page. (2018). Development and Evaluation of a Transfusion Medicine Genome Wide Genotyping Array. *Transfusion, in press*.

Uncategorized References

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics, 76*(1), 7.20. 21-27.20. 41.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., . . . Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage, 166*, 400-424. doi:10.1016/j.neuroimage.2017.10.034

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *science, 322*(5903), 881-888.

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., . . . Malik, R. (2018). Analysis of shared heritability in common disorders of the brain. *Science, 360*(6395).

APA. (2000). *Diagnostic and Statistic Manual of Mental Disorders, 4th ed Text Revision DSM-IV-TR*. Washington, DC: American Psychiatric Press

APA. (2013). *Diagnostic and Statistic Manual of Mental Disorders, 5th ed Text Revision DSM-V*. Washington, DC: American Psychiatric Press

Arnold, P. D., Askland, K. D., Barlassina, C., Bellodi, L., Bienvenu, O., Black, D., . . . Camarena, B. (2018). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Molecular psychiatry, 23*(5), 1181-1181.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., . . . Kostadima, M. A. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell, 167*(5), 1415-1429. e1419. doi:10.1016/j.cell.2016.10.042

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., . . . Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science, 297*(5583), 1003-1007. Retrieved from http://science.sciencemag.org/content/sci/297/5583/1003.full.pdf

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome research, 11*(6), 1005-1017. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC311093/pdf/X24PU.pdf

Benonisdottir, S., Oddsson, A., Helgason, A., Kristjansson, R. P., Sveinbjornsson, G., Oskarsdottir, A., . . . Sulem, G. (2016). Epigenetic and genetic components of height regulation. *Nature Communications, 7*, 13490. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5116096/pdf/ncomms13490.pdf

Bertelsen, B., Stefánsson, H., Jensen, L. R., Melchior, L., Debes, N. M., Groth, C., . . . Tarnok, Z. (2016). Association of AADAC deletion and Gilles de la tourette syndrome in a large european cohort. *Biological psychiatry, 79*(5), 383-391.

Bjornsdottir, G., Ivarsdottir, E. V., Bjarnadottir, K., Benonisdottir, S., Gylfadottir, S. S., Arnadottir, G. A., . . . Jonasdottir, A. (2019). A PRPH splice-donor variant associates with reduced sural nerve amplitude and risk of peripheral neuropathy. *Nature communications, 10*(1), 1-10.

Brander, G., Rydell, M., Kuja-Halkola, R., de la Cruz, L. F., Lichtenstein, P., Serlachius, E., . . . Larsson, H. (2018). Perinatal risk factors in Tourette's and chronic tic disorders: a total population sibling comparison study. *Molecular psychiatry, 23*(5), 1189-1197.

Breen, G., Li, Q., Roth, B. L., O'Donnell, P., Didriksen, M., Dolmetsch, R., . . . Huebel, C. (2016). Translating genome-wide association findings into new therapeutics for psychiatry. *Nature neuroscience, 19*(11), 1392-1396.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics, 12*(10), 703-714.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., . . . Robinson, E. B. (2015). An atlas of genetic correlations across human diseases and traits. *Nature genetics, 47*(11), 1236.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., . . . Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics, 47*(3), 291-295.

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., & Howson, J. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature communications, 11*(1), 1-11.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209. doi:10.1038/s41586-018-0579-z

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209.

Calafato, M. S., Thygesen, J. H., Ranlund, S., Zartaloudi, E., Cahn, W., Crespo-Facorro, B., . . . Hall, M.-H. (2018). Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders. *The British Journal of Psychiatry, 213*(3), 535-541.

Cao, Q., Zang, Y., Sun, L., Sui, M., Long, X., Zou, Q., & Wang, Y. (2006). Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study. *Neuroreport, 17*(10), 1033-1036.

Cavanna, A. E., Servo, S., Monaco, F., & Robertson, M. M. (2009). The behavioral spectrum of Gilles de la Tourette syndrome. *J Neuropsychiatry Clin Neurosci, 21*(1), 13-23. doi:10.1176/appi.neuropsych.21.1.13

Cheung, V. G., & Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews Genetics, 10*(9), 595-604.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Campbell, P. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712.

Consortium, S. W. G. o. t. P. G. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427.

Crouch, D. J. M., & Bodmer, W. F. (2020). Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc Natl Acad Sci U S A, 117*(32), 18924-18933. doi:10.1073/pnas.2005634117

Darrow, S. M., Hirschtritt, M. E., Davis, L. K., Illmann, C., Osiecki, L., Grados, M., . . . Pauls, D. (2017). Identification of two heritable cross-disorder endophenotypes for Tourette syndrome. *American Journal of Psychiatry, 174*(4), 387-396.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., . . . Bækvad-Hansen, M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics, 51*(1), 63-75.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., . . . Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet, 51*(1), 63-75. doi:10.1038/s41588-018-0269-7

Didriksen, M., Nawaz, M. S., Dowsett, J., Bell, S., Erikstrup, C., Pedersen, O. B., . . . Stefansson, K. (2020). Large genome-wide association study identifies three novel risk variants for restless legs syndrome. *Commun Biol, 3*(1), 703. doi:10.1038/s42003-020-01430-1

Didriksen, M., Rigas, A. S., Allen, R. P., Burchell, B. J., Di Angelantonio, E., Nielsen, M. H., . . . Pedersen, O. B. (2017). Prevalence of restless legs syndrome and associated factors in an otherwise healthy population: results from the Danish Blood Donor Study. *Sleep medicine, 36*, 55-61.

Didriksen, M., Rigas, A. S., Allen, R. P., Burchell, B. J., Di Angelantonio, E., Nielsen, M. H., . . . Ullum, H. (2017). Prevalence of restless legs syndrome and associated factors in an otherwise healthy population: results from the Danish Blood Donor Study. *Sleep Med, 36*, 55-61. doi:10.1016/j.sleep.2017.04.014

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., . . . Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic acids research, 36*(19), e126-e126. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2577347/pdf/gkn556.pdf

Emdin, C. A., Khera, A. V., & Kathiresan, S. (2017). Mendelian randomization. *Jama, 318*(19), 1925-1926.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., . . . Gunnarsdottir, S. (2008). Genetics of gene expression and its effect on disease. *Nature, 452*(7186), 423-428.

Eriksson, J., Haring, R., Grarup, N., Vandenput, L., Wallaschofski, H., Lorentzen, E., . . . Nauck, M. (2017). Causal relationship between obesity and serum testosterone status in men: A bi-directional mendelian randomization analysis. *PloS one, 12*(4), e0176277.

Estruch, M., Bancells, C., Beloki, L., Sanchez-Quesada, J. L., Ordonez-Llanos, J., & Benitez, S. (2013). CD14 and TLR4 mediate cytokine release promoted by electronegative LDL in monocytes. *Atherosclerosis, 229*(2), 356-362. doi:10.1016/j.atherosclerosis.2013.05.011

Evangelou, E., & Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics, 14*(6), 379-389.

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., . . . Karaman, I. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics, 50*(10), 1412-1425.

Fernandez, T. V., Sanders, S. J., Yurkiewicz, I. R., Ercan-Sencicek, A. G., Kim, Y.-S., Fishman, D. O., . . . Ho, W. S. (2012). Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. *Biological psychiatry, 71*(5), 392-402.

Forde, N. J., Kanaan, A. S., Widomska, J., Padmanabhuni, S. S., Nespoli, E., Alexander, J., . . . Nawaz, M. S. (2016). TS-EUROTRAIN: a European-wide investigation and training network on the etiology and pathophysiology of Gilles de la Tourette syndrome. *Frontiers in neuroscience, 10*, 384.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., . . . Hurles, M. E. (2006). Copy number variation: new insights in genome diversity. *Genome research, 16*(8), 949-961.

Gisladottir, R. S., Ivarsdottir, E. V., Helgason, A., Jonsson, L., Hannesdottir, N. K., Rutsdottir, G., . . . Norddahl, G. L. (2020). Sequence variants in TAAR5 and other loci affect human odor perception and naming. *Current Biology, 30*(23), 4643-4653. e4643.

Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., . . . Palotie, A. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat Genet, 48*(8), 856-866. doi:10.1038/ng.3598

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., . . . Helgadottir, A. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics, 38*(3), 320-323.

Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., . . . Enhancing NeuroImaging Genetics through Meta-Analysis Consortium -Genetics working, g. (2020). The genetic architecture of the human cerebral cortex. *Science, 367*(6484). doi:10.1126/science.aay6690

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., . . . Borglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet, 51*(3), 431-444. doi:10.1038/s41588-019-0344-8

Gudbjartsson, D. F., Arnar, D. O., Helgadottir, A., Gretarsdottir, S., Holm, H., Sigurdsson, A., . . . Kristjansson, K. (2007). Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature, 448*(7151), 353-357.

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., . . . Hjartarson, E. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics, 47*(5), 435-444.

Gudbjartsson, D. F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S. A., Zink, F., . . . Hjartarson, E. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. *Scientific data, 2*.

Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J. T., Manolescu, A., Gudbjartsson, D., . . . Blondal, T. (2008). Common sequence variants on 2p15 and Xp11. 22 confer susceptibility to prostate cancer. *Nature genetics, 40*(3), 281-283.

Gudmundsson, J., Thorleifsson, G., Sigurdsson, J. K., Stefansdottir, L., Jonasson, J. G., Gudjonsson, S. A., . . . Halldorsson, G. H. (2017). A genome-wide association study yields five novel thyroid cancer risk loci. *Nature communications, 8*(1), 1-8.

Gudmundsson, O. O., Walters, G. B., Ingason, A., Johansson, S., Zayats, T., Athanasiu, L., . . . Jonsson, G. F. (2019). Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. *Translational psychiatry, 9*(1), 1-9.

Gul, A., Hassan, M. J., Mahmood, S., Chen, W., Rahmani, S., Naseer, M. I., . . . Ansar, M. (2006). Genetic studies of autosomal recessive primary microcephaly in 33 Pakistani families: novel sequence variants in ASPM gene. *Neurogenetics, 7*(2), 105-110.

Haworth, S., Shapland, C. Y., Hayward, C., Prins, B. P., Felix, J. F., Medina-Gomez, C., . . . Vrijheid, M. (2019). Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. *Nature communications, 10*(1), 1-16.

Haworth, S., Shapland, C. Y., Hayward, C., Prins, B. P., Felix, J. F., Medina-Gomez, C., . . . St Pourcain, B. (2019). Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. *Nat Commun, 10*(1), 357. doi:10.1038/s41467-018-07863-x

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21*(11), 1539-1558.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362-9367.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics, 6*(2), 95-108.

Hirschtritt, M. E., Bloch, M. H., & Mathews, C. A. (2017). Obsessive-compulsive disorder: advances in diagnosis and treatment. *Jama, 317*(13), 1358-1367.

Hirschtritt, M. E., Lee, P. C., Pauls, D. L., Dion, Y., Grados, M. A., Illmann, C., . . . Lyon, G. J. (2015). Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. *JAMA psychiatry, 72*(4), 325-333. Retrieved from http://archpsyc.jamanetwork.com/data/journals/psych/933641/yoi140111.pdf

Holmes, M. V., Asselbergs, F. W., Palmer, T. M., Drenos, F., Lanktree, M. B., Nelson, C. P., . . . Swerdlow, D. I. (2015). Mendelian randomization of blood lipids for coronary heart disease. *European heart journal, 36*(9), 539-550.

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., . . . Wigmore, E. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience, 22*(3), 343-352.

Hshieh, T. T., Fox, M. L., Kosar, C. M., Cavallari, M., Guttmann, C. R., Alsop, D., . . . Inouye, S. K. (2016). Head circumference as a useful surrogate for intracranial volume in older adults. *Int Psychogeriatr, 28*(1), 157-162. doi:10.1017/S104161021500037X

Hsu, Y. H., Estrada, K., Evangelou, E., Ackert-Bicknell, C., Akesson, K., Beck, T., . . . Cauley, J. (2019). Meta-analysis of genomewide association studies reveals genetic variants for hip bone geometry. *Journal of Bone and Mineral Research, 34*(7), 1284-1296.

Hurst, J. A., Jenkins, D., Vasudevan, P. C., Kirchhoff, M., Skovby, F., Rieubland, C., . . . Wilkie, A. O. (2011). Metopic and sagittal synostosis in Greig cephalopolysyndactyly syndrome: five cases with intragenic mutations or complete deletions of GLI3. *Eur J Hum Genet, 19*(7), 757-762. doi:10.1038/ejhg.2011.13

Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietiläinen, O., . . . Muglia, P. (2011). Copy number variations of chromosome 16p13. 1 region associated with schizophrenia. *Molecular psychiatry, 16*(1), 17-25.

International League Against Epilepsy Consortium on Complex, E. (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat Commun, 9*(1), 5269. doi:10.1038/s41467-018-07524-z

International Obsessive Compulsive Disorder Foundation Genetics, C., & Studies, O. C. D. C. G. A. (2018). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psychiatry, 23*(5), 1181-1188. doi:10.1038/mp.2017.154

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., . . . Athanasiu, L. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics, 51*(3), 404-413.

Jansen, P. R., Nagel, M., Watanabe, K., Wei, Y., Savage, J. E., de Leeuw, C. A., . . . Posthuma, D. (2019). GWAS of brain volume on 54,407 individuals and cross-trait analysis with intelligence identifies shared genomic loci and genes. *BioRxiv*, 613489.

Jansen, P. R., Nagel, M., Watanabe, K., Wei, Y., Savage, J. E., de Leeuw, C. A., . . . Posthuma, D. (2020). Genome-wide meta-analysis of brain volume identifies genomic loci and genes shared with intelligence. *Nat Commun, 11*(1), 5606. doi:10.1038/s41467-020-19378-5

Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., . . . Posthuma, D. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet, 51*(3), 394-403. doi:10.1038/s41588-018-0333-3

Jimenez-Sanchez, G., Childs, B., & Valle, D. (2001). Human disease genes. *Nature, 409*(6822), 853-855.

Jónsson, B. A., Bjornsdottir, G., Thorgeirsson, T., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D., . . . Ulfarsson, M. (2019). Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications, 10*(1), 1-10.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., . . . Gudjonsson, S. A. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature, 549*(7673), 519-522.

Jonsson, L., Magnusson, T. E., Thordarson, A., Jonsson, T., Geller, F., Feenstra, B., . . . Stefansson, K. (2018). Rare and Common Variants Conferring Risk of Tooth Agenesis. *J Dent Res, 97*(5), 515-522. doi:10.1177/0022034517750109

Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., . . . Maloney, J. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature, 488*(7409), 96-99.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W.-m., Huang, J., Liu, G., . . . Zhang, J. (2003). Large-scale genotyping of complex DNA. *Nature biotechnology, 21*(10), 1233-1237.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry, 62*(6), 593-602.

Khachatryan, S. G., Ferri, R., Fulda, S., Garcia-Borreguero, D., Manconi, M., Muntean, M. L., & Stefani, A. (2022). Restless legs syndrome: Over 50 years of E uropean contribution. *Journal of sleep research, 31*(4), e13632.

Khalifa, N., & von Knorring, A. L. (2003). Prevalence of tic disorders and Tourette syndrome in a Swedish school population. *Dev Med Child Neurol, 45*(5), 315-319. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12729145

http://onlinelibrary.wiley.com/store/10.1111/j.1469-8749.2003.tb00402.x/asset/j.1469-8749.2003.tb00402.x.pdf?v=1&t=j655dsfc&s=726025c629291fc2b8cb33ebff4e938cb92fe4a8

Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., . . . Ropers, H.-H. (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Human molecular genetics, 17*(3), 458-465.

Kirov, G., Rees, E., Walters, J. T., Escott-Price, V., Georgieva, L., Richards, A. L., . . . Moran, J. L. (2014). The penetrance of copy number variations for schizophrenia and developmental delay. *Biological psychiatry, 75*(5), 378-385.

Klein, M., Walters, R. K., Demontis, D., Stein, J. L., Hibar, D. P., Adams, H. H., . . . Sonuga-Barke, E. (2019). Genetic markers of ADHD-related variations in intracranial volume. *American Journal of Psychiatry, 176*(3), 228-238.

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

Kong, A., Frigge, M. L., Thorleifsson, G., Stefansson, H., Young, A. I., Zink, F., . . . Masson, G. (2017). Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences, 114*(5), E727-E732.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., . . . Masson, G. (2002). A high-resolution recombination map of the human genome. *Nature genetics, 31*(3), 241-247.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., . . . Rafnar, T. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics, 40*(9), 1068-1075. Retrieved from http://www.nature.com/ng/journal/v40/n9/pdf/ng.216.pdf

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., . . . Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet, 40*(9), 1068-1075. doi:10.1038/ng.216

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., . . . Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science, 359*(6374), 424-428. doi:10.1126/science.aan6877

Laansma, M. A., Bright, J. K., Al-Bachari, S., Anderson, T. J., Ard, T., Assogna, F., . . . Study, E. N.-P. s. (2021). International Multicenter Analysis of Brain Structure Across Clinical Stages of Parkinson's Disease. *Mov Disord, 36*(11), 2583-2594. doi:10.1002/mds.28706

Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., . . . Brown, B. C. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature genetics, 51*(12), 1670-1678.

Lam, M., Chen, C. Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., . . . Huang, H. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet, 51*(12), 1670-1678. doi:10.1038/s41588-019-0512-x

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . FitzHugh, W. (2001). Initial sequencing and analysis of the human genome.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., . . . Linnér, R. K. (2018a). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature genetics, 50*(8), 1112.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., . . . Linnér, R. K. (2018b). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics, 50*(8), 1112-1121.

Lee, P. H., Feng, Y.-C. A., & Smoller, J. W. (2021). Pleiotropy and cross-disorder genetics among psychiatric disorders. *Biological psychiatry, 89*(1), 20-31.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Cummings, B. B. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291.

Lesperance, P., Djerroud, N., Diaz Anzaldua, A., Rouleau, G., Chouinard, S., & Richer, F. (2004). Restless legs in Tourette syndrome. *Movement disorders, 19*(9), 1084-1087.

Levey, D. F., Stein, M. B., Wendt, F. R., Pathak, G. A., Zhou, H., Aslan, M., . . . McIntosh, A. M. (2020). GWAS of Depression Phenotypes in the Million Veteran Program and Meta-analysis in More than 1.2 Million Participants Yields 178 Independent Risk Loci. *medRxiv*.

Levy, A. M., Paschou, P., & Tümer, Z. (2021). Candidate Genes and Pathways Associated with Gilles de la Tourette Syndrome—Where Are We? *Genes, 12*(9), 1321.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Tian, C. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics, 51*(2), 237-244.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet, 51*(2), 237-244. doi:10.1038/s41588-018-0307-5

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., . . . Kauppi, K. (2017). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature genetics, 49*(1), 152-156.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., . . . Berger, B. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics, 47*(3), 284.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., . . . Grarup, N. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics, 50*(11), 1505-1513.

Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell, 148*(6), 1223-1241.

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., . . . Gretarsdottir, S. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics, 50*(4), 524-537.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics, 39*(7), 906-913.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet, 39*(7), 906-913. doi:10.1038/ng2088

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., . . . Ren, Y. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics, 82*(2), 477-488.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Brody, J. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science, 337*(6099), 1190-1195.

McGrath, L. M., Yu, D., Marshall, C., Davis, L. K., Thiruvahindrapuram, B., Li, B., . . . Schroeder, F. A. (2014). Copy number variation in obsessive-compulsive disorder and tourette syndrome: a cross-disorder study. *Journal of the American Academy of Child & Adolescent Psychiatry, 53*(8), 910-919.

Miller, O. J., & Therman, E. (2011). *Human chromosomes*: Springer Science & Business Media.

Mills, M. C., Tropf, F. C., Brazel, D. M., Van Zuydam, N., Vaez, A., Pers, T. H., . . . Den Hoed, M. (2020). Identification of 370 loci for age at onset of sexual and reproductive behaviour, highlighting common aetiology with reproductive biology, externalizing behaviour and longevity. *BioRxiv*.

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., & He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 1-7.

Morrow, E. M. (2010). Genomic copy number variation in disorders of cognitive development. *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(11), 1091-1104.

Mort, M., Ivanov, D., Cooper, D. N., & Chuzhanova, N. A. (2008). A meta-analysis of nonsense mutations causing human genetic disease. *Human mutation, 29*(8), 1037-1047.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification, 31*(3), 274-295.

Nag, A., Bochukova, E. G., Kremeyer, B., Campbell, D. D., Muller, H., Valencia-Duarte, A. V., . . . Cuartas, M. (2013). CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. *Plos one, 8*(3), e59061.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., De Leeuw, C. A., Bryois, J., . . . Muñoz-Manchado, A. B. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics, 50*(7), 920-927.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., Bryois, J., . . . Posthuma, D. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet, 50*(7), 920-927. doi:10.1038/s41588-018-0151-7

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., . . . Xue, A. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology, 18*(12), 1091-1102.

Neumann, W. J., Schroll, H., de Almeida Marcelino, A. L., Horn, A., Ewert, S., Irmen, F., . . . Kuhn, A. A. (2018). Functional segregation of basal ganglia pathways in Parkinson's disease. *Brain, 141*(9), 2655-2669. doi:10.1093/brain/awy206

Nicholas, A. K., Khurshid, M., Désir, J., Carvalho, O. P., Cox, J. J., Thornton, G., . . . Verloes, A. (2010). WDR62 is associated with the spindle pole and is mutated in human microcephaly. *Nature genetics, 42*(11), 1010-1014.

Nielsen, J. B., Thorolfsdottir, R. B., Fritsche, L. G., Zhou, W., Skov, M. W., Graham, S. E., . . . Sveinbjornsson, G. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature genetics, 50*(9), 1234-1239.

Noonan, S. K., Haist, F., & Müller, R.-A. (2009). Aberrant functional connectivity in autism: evidence from low-frequency BOLD signal fluctuations. *Brain research, 1262*, 48-63.

Norland, K., Sveinbjornsson, G., Thorolfsdottir, R. B., Davidsson, O. B., Tragante, V., Rajamani, S., . . . Asselbergs, F. W. (2019). Sequence variants with large effects on cardiac electrophysiology and disease. *Nature communications, 10*(1), 1-10.

Ohi, K., Ochi, R., Noda, Y., Wada, M., Sugiyama, S., Nishi, A., . . . Nakajima, S. (2021). Polygenic risk scores for major psychiatric and neurodevelopmental disorders contribute to sleep disturbance in childhood: Adolescent Brain Cognitive Development (ABCD) Study. *Translational psychiatry, 11*(1), 1-11.

Olafsdottir, T., Stacey, S. N., Sveinbjornsson, G., Thorleifsson, G., Norland, K., Sigurgeirsson, B., . . . Sarin, K. Y. (2021). Loss-of-Function Variants in the Tumor-Suppressor Gene PTPN14 Confer Increased Cancer Risk. *Cancer Research*.

Olafsdottir, T., Stacey, S. N., Sveinbjornsson, G., Thorleifsson, G., Norland, K., Sigurgeirsson, B., . . . Stefansson, K. (2021). Loss-of-Function Variants in the Tumor-Suppressor Gene PTPN14 Confer Increased Cancer Risk. *Cancer Res, 81*(8), 1954-1964. doi:10.1158/0008-5472.CAN-20-3065

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics, 5*(5), 389-396.

Paschou, P., Jin, Y., Müller-Vahl, K., Möller, H. E., Rizzo, R., Hoekstra, P. J., . . . Hartmann, A. (2022). Enhancing neuroimaging genetics through meta-analysis for Tourette syndrome (ENIGMA-TS): A worldwide platform for collaboration. *Frontiers in Psychiatry*, 1763.

Pedrazzini, B., Waldvogel, S., Vaucher, P., Cornuz, J., Heinzer, R., Tissot, J. D., & Favrat, B. (2014). Prevalence of restless legs syndrome in female blood donors 1 week after blood donation. *Vox sanguinis, 107*(1), 44-49.

Piacentini, J., & Chang, S. (2005). Habit reversal training for tic disorders in children and adolescents. *Behavior Modification, 29*(6), 803-822.

Pierce, J. E., & Peron, J. (2020). The basal ganglia and the cerebellum in human emotion. *Soc Cogn Affect Neurosci, 15*(5), 599-613. doi:10.1093/scan/nsaa076

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., . . . Mills, R. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology, 29*(6), 512-520. Retrieved from http://www.nature.com/nbt/journal/v29/n6/pdf/nbt.1852.pdf

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., . . . Wang, Z. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics, 94*(5), 677-694.

Power, R. A., Steinberg, S., Bjornsdottir, G., Rietveld, C. A., Abdellaoui, A., Nivard, M. M., . . . Willemsen, G. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature neuroscience, 18*(7), 953-955.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., . . . Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature methods, 5*(12), 1005-1010.

Rees, E., Kirov, G., Sanders, A., Walters, J. T. R., Chambert, K., Shi, J., . . . Green, E. K. (2014). Evidence that duplications of 22q11. 2 protect against schizophrenia. *Molecular Psychiatry, 19*(1), 37-40.

Rees, E., O'Donovan, M. C., & Owen, M. J. (2015). Genetics of schizophrenia. *Current Opinion in Behavioral Sciences, 2*, 8-14.

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *TRENDS in Genetics, 17*(9), 502-510.

Ripke, S., Walters, J. T., O'Donovan, M. C., & Consortium, S. W. G. o. t. P. G. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *MedRxiv*.

Ritchie, G. R., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature methods, 11*(3), 294-296.

Robertson, M. M., Eapen, V., & Cavanna, A. E. (2009). The international prevalence, epidemiology, and clinical phenomenology of Tourette syndrome: a cross-cultural perspective. *J Psychosom Res, 67*(6), 475-483. doi:10.1016/j.jpsychores.2009.07.010

Robertson, M. M., Eapen, V., Singer, H. S., Martino, D., Scharf, J. M., Paschou, P., . . . Mathews, C. A. (2017). Gilles de la Tourette syndrome. *Nature reviews Disease primers, 3*(1), 1-20.

Saevarsdottir, S., Olafsdottir, T. A., Ivarsdottir, E. V., Halldorsson, G. H., Gunnarsdottir, K., Sigurdsson, A., . . . Lund, S. H. (2020). FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature*, 1-5.

Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N. F., Guyatt, A. L., Jackson, V. E., . . . Kim, D. K. (2019). Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nature genetics, 51*(3), 494-505.

Santos, R. L. P., Wajid, M., Khan, M. N., McArthur, N., Pham, T. L., Bhatti, A., . . . Yan, K. (2005). Novel sequence variants in the TMC1 gene in Pakistani families with autosomal recessive hearing impairment. *Human mutation, 26*(4), 396-396.

Satizabal, C. L., Adams, H. H., Hibar, D. P., White, C. C., Knol, M. J., Stein, J. L., . . . Roshchupkin, G. V. (2019). Genetic architecture of subcortical brain structures in 38,851 individuals. *Nature genetics, 51*(11), 1624-1636.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., . . . Coleman, J. R. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics, 50*(7), 912-919.

Scharf, J. M., Miller, L. L., Gauvin, C. A., Alabiso, J., Mathews, C. A., & Ben-Shlomo, Y. (2014). Population prevalence of Tourette syndrome: A systematic review and meta-analysis. *Mov Disord*. doi:10.1002/mds.26089

Scharf, J. M., Miller, L. L., Mathews, C. A., & Ben-Shlomo, Y. (2012). Prevalence of Tourette syndrome and chronic tics in the population-based Avon longitudinal study of parents and children cohort. *J Am Acad Child Adolesc Psychiatry, 51*(2), 192-201 e195. doi:10.1016/j.jaac.2011.11.004

Schneider, M., Debbané, M., Bassett, A. S., Chow, E. W., Fung, W. L. A., Van Den Bree, M. B., . . . Kates, W. R. (2014). Psychiatric disorders from childhood to adulthood in 22q11. 2 deletion syndrome: results from the International Consortium on Brain and Behavior in 22q11. 2 Deletion Syndrome. *American Journal of Psychiatry, 171*(6), 627-639.

Schormair, B., Zhao, C., Bell, S., Tilch, E., Salminen, A. V., Pütz, B., . . . Poewe, W. (2017). Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. *The Lancet Neurology, 16*(11), 898-907.

Schormair, B., Zhao, C., Bell, S., Tilch, E., Salminen, A. V., Putz, B., . . . Winkelmann, J. (2017). Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. *Lancet Neurol, 16*(11), 898-907. doi:10.1016/s1474-4422(17)30327-7

Senanayake, K., Krashin, D., & Murinova, N. (2020). Comorbidities of Patients with Restless Legs Syndrome (4994). In: AAN Enterprises.

Shameer, K., Tripathi, L. P., Kalari, K. R., Dudley, J. T., & Sowdhamini, R. (2016). Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Briefings in bioinformatics, 17*(5), 841-862.

Shrine, N., Guyatt, A. L., Erzurumluoglu, A. M., Jackson, V. E., Hobbs, B. D., Melbourne, C. A., . . . Sakornsakolpat, P. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nature genetics, 51*(3), 481-493.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research, 40*(W1), W452-W457.

Smit, D. J., Cath, D., Zilhão, N. R., Ip, H. F., Denys, D., den Braber, A., . . . Boomsma, D. I. (2020). Genetic meta-analysis of obsessive–compulsive disorder and self-report compulsive symptoms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 183*(4), 208-216.

Sonderby, I. E., Gustafsson, O., Doan, N. T., Hibar, D. P., Martin-Brevet, S., Abdellaoui, A., . . . p11.2 European Consortium, f. t. E.-C. N. V. w. g. (2020). Dose response of the 16p11.2 distal copy number variant on intracranial volume and basal ganglia. *Mol Psychiatry, 25*(3), 584-602. doi:10.1038/s41380-018-0118-1

Stacey, S. N., Sulem, P., Johannsson, O. T., Helgason, A., Gudmundsson, J., Kostic, J. P., . . . Hrafnkelsson, J. (2006). The BARD1 Cys557Ser variant and breast cancer risk in Iceland. *PLoS Med, 3*(7), e217.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., . . . Gaspar, H. A. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics, 51*(5), 793-803.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., . . . Bipolar Disorder Working Group of the Psychiatric Genomics, C. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet, 51*(5), 793-803. doi:10.1038/s41588-019-0397-8

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., . . . Gudnadottir, V. G. (2005). A common inversion under selection in Europeans. *Nature genetics, 37*(2), 129-137.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., . . . Doyle, O. M. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature, 505*(7483), 361-366.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., . . . Buizer-Voskamp, J. E. (2008). Large recurrent microdeletions associated with schizophrenia. *nature, 455*(7210), 232-236.

Stefansson, H., Rye, D. B., Hicks, A., Petursson, H., Ingason, A., Thorgeirsson, T. E., . . . Eiriksdottir, I. (2007). A genetic risk factor for periodic limb movements in sleep. *New England journal of medicine, 357*(7), 639-647.

Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., . . . Unnsteinsdottir, U. (2015). Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature genetics, 47*(5), 445-447.

Steinthorsdottir, V., Thorleifsson, G., Aradottir, K., Feenstra, B., Sigurdsson, A., Stefansdottir, L., . . . Nielsen, N. M. (2016). Common variants upstream of KDR encoding VEGFR2 and in TTC39B associate with endometriosis. *Nature communications, 7*.

Stram, D. O. (2017). Multi-SNP haplotype analysis methods for association analysis. In *Statistical Human Genetics* (pp. 485-504): Springer.

Styrkarsdottir, U., Helgason, H., Sigurdsson, A., Norddahl, G. L., Agustsdottir, A. B., Reynard, L. N., . . . Magnusdottir, A. (2017). Whole-genome sequencing identifies rare genotypes in COMP and CHADL associated with high risk of hip osteoarthritis. *Nature genetics, 49*(5), 801-805.

Styrkarsdottir, U., Stefansson, O. A., Gunnarsdottir, K., Thorleifsson, G., Lund, S. H., Stefansdottir, L., . . . Halldorsson, G. H. (2019). GWAS of bone size yields twelve loci that also affect height, BMD, osteoarthritis or fractures. *Nature communications, 10*(1), 1-13.

Sundaram, S. K., Huq, A. M., Wilson, B. J., & Chugani, H. T. (2010). Tourette syndrome is associated with recurrent exonic copy number variants. *Neurology, 74*(20), 1583-1590.

Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., . . . Sulem, P. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature genetics*.

Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics, 3*(5), 391-397.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics, 20*(8), 467-484.

Thomsen, S. K., & Gloyn, A. L. (2017). Human genetics as a model for target validation: finding new therapies for diabetes. *Diabetologia, 60*(6), 960-970.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., . . . Ingason, A. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature, 452*(7187), 638-642.

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., . . . Walter, S. (2010). Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking behavior. *Nature genetics, 42*(5), 448-453.

Tian, L., Jiang, T., Wang, Y., Zang, Y., He, Y., Liang, M., . . . Peng, M. (2006). Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neuroscience letters, 400*(1-2), 39-43.

Trotti, L. M., Bliwise, D. L., Greer, S. A., Sigurdsson, A. P., Gudmundsdóttir, G. B., Wessel, T., . . . Sigmundsson, T. (2009). Correlates of PLMs variability over multiple nights and impact upon RLS diagnosis. *Sleep medicine, 10*(6), 668-671.

Usuba, R., Pauty, J., Soncin, F., & Matsunaga, Y. T. (2019). EGFL7 regulates sprouting angiogenesis and endothelial integrity in a human blood vessel model. *Biomaterials, 197*, 305-316.

Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J. S., & Kutalik, Z. (2013). The growing importance of CNVs: new insights for detection and clinical interpretation. *Frontiers in genetics, 4*, 92. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3667386/pdf/fgene-04-00092.pdf

Van de Griendt, J., Verdellen, C., Van Dijk, M., & Verbraak, M. (2013). Behavioural treatment of tics: habit reversal and exposure with response prevention. *Neuroscience & Biobehavioral Reviews, 37*(6), 1172-1177.

van der Harst, P., & Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research, 122*(3), 433-443.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.), 25*(3), 427.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Holt, R. A. (2001). The sequence of the human genome. *science, 291*(5507), 1304-1351.

Vink, J. M., Treur, J. L., Pasman, J. A., & Schellekens, A. (2020). Investigating genetic correlation and causality between nicotine dependence and ADHD in a broader psychiatric context. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics, 101*(1), 5-22.

Walters, G. B., Gustafsson, O., Sveinbjornsson, G., Eiriksdottir, V. K., Agustsdottir, A. B., Jonsdottir, G. A., . . . Stefansson, K. (2018). MAP1B mutations cause intellectual disability and extensive white matter deficit. *Nat Commun, 9*(1), 3456. doi:10.1038/s41467-018-05595-6

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., . . . Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research, 17*(11), 1665-1674. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2045149/pdf/1665.pdf

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., . . . Sleiman, P. M. (2009). Common genetic variants on 5p14. 1 associate with autism spectrum disorders. *Nature, 459*(7246), 528-533.

Warrington, N. M., Beaumont, R. N., Horikoshi, M., Day, F. R., Helgeland, O., Laurin, C., . . . Freathy, R. M. (2019). Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet, 51*(5), 804-814. doi:10.1038/s41588-019-0403-1

Watanabe, K., Jansen, P. R., Savage, J. E., Nandakumar, P., Wang, X., Hinds, D. A., . . . Stein, M. (2020). Genome-wide meta-analysis of insomnia in over 2.3 million individuals indicates involvement of specific biological pathways through gene-prioritization. *medRxiv*.

Watanabe, K., Stringer, S., Frei, O., Umicevic Mirkov, M., de Leeuw, C., Polderman, T. J. C., . . . Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet, 51*(9), 1339-1348. doi:10.1038/s41588-019-0481-0

Watanabe, K., Taskesen, E., Van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature communications, 8*(1), 1-11.

Watson, H. J., Yilmaz, Z., Thornton, L. M., Hubel, C., Coleman, J. R. I., Gaspar, H. A., . . . Bulik, C. M. (2019). Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet, 51*(8), 1207-1214. doi:10.1038/s41588-019-0439-2

Weng, S.-J., Wiggins, J. L., Peltier, S. J., Carrasco, M., Risi, S., Lord, C., & Monk, C. S. (2010). Alterations of resting state functional connectivity in the default network in adolescents with autism spectrum disorders. *Brain research, 1313*, 202-214.

WHO. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines.* Geneva: World Health Organization.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191.

Williams, N. M., Zaharieva, I., Martin, A., Langley, K., Mantripragada, K., Fossdal, R., . . . Gudmundsson, O. O. (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet, 376*(9750), 1401-1408.

Winter-Jensen, M., Afzal, S., Jess, T., Nordestgaard, B. G., & Allin, K. H. (2020). Body mass index and risk of infections: a Mendelian randomization study of 101,447 individuals. *European journal of epidemiology, 35*(4), 347-354.

Woods, D. W., & Thomsen, P. H. (2014). Tourette and tic disorders in ICD-11: standing at the diagnostic crossroads. *Rev Bras Psiquiatr, 36 Suppl 1*, 51-58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25388612

http://www.scielo.br/pdf/rbp/v36s1/1516-4446-rbp-2013-36-S1-S51.pdf

Xu, K., Li, B., McGinnis, K. A., Vickers-Smith, R., Dao, C., Sun, N., . . . Gelernter, J. (2020). Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nature communications, 11*(1), 1-11.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., . . . Loos, R. J. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics, 44*(4), 369-375.

Yang, Z., Wu, H., Lee, P. H., Tsetsos, F., Davis, L. K., Yu, D., . . . Barta, C. (2021). Investigating shared genetic basis across tourette syndrome and comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum. *Biological psychiatry, 90*(5), 317-327.

Yeh, P., Walters, A. S., & Tsuang, J. W. (2012). Restless legs syndrome: a comprehensive overview on its epidemiology, risk factors, and treatment. *Sleep and Breathing, 16*(4), 987-1007.

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., . . . Visscher, P. M. (2018). Meta-analysis of genome-wide association studies for height and body mass index in~ 700000 individuals of European ancestry. *Human molecular genetics, 27*(20), 3641-3649.

Yu, C. H., Pal, L. R., & Moult, J. (2016). Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease. *OMICS, 20*(7), 400-414. doi:10.1089/omi.2016.0063

Yu, D., Sul, J. H., Tsetsos, F., Nawaz, M. S., Huang, A. Y., Zelaya, I., . . . Hirschtritt, M. E. (2019). Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies. *American Journal of Psychiatry, 176*(3), 217-227.

Zhang, F., Baranova, A., Zhou, C., Cao, H., Chen, J., Zhang, X., & Xu, M. (2021). Causal influences of neuroticism on mental health and cardiovascular disease. *Human Genetics*, 1-15.

Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., & Wang, H. (2014). Evaluation of copy number variation detection for a SNP array platform. *BMC bioinformatics, 15*(1), 1. Retrieved from https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-15-1?site=bmcbioinformatics.biomedcentral.com

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., . . . Zhu, Z. (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature genetics, 51*(11), 1637-1644.

Zink, F., Stacey, S. N., Norddahl, G. L., Frigge, M. L., Magnusson, O. T., Jonsdottir, I., . . . Gudmundsson, J. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood, 130*(6), 742-752.

Original publications

Paper I

Paper II

Paper III

Paper IV

Paper V

Appendix

Appendix 1 – Brief TS/TD screening questionnaire (informally translated from Icelandic) based on ICD-10 and DSM-IV-TR diagnostic criteria.

1. Have you ever been diagnosed with a Tic disorder?

2. Have you ever been diagnosed with Tourette syndrome?

3. Have you ever had any involuntary tics that started before you were 18?

4. If yes to question 3, do/did you have episodes of

4.1. repeated clicking of finger-joints

4.2. repeated eye-blinking

4.3. repeated facial grimaces

4.4. repeated head jerks

4.5. repeated mouth-twitches or mouth-grimaces

4.6. repeated movements (or tensing/stretching) of your legs, feet, or toes

4.7. repeated nose-twitching or sniffing

4.8. repeated upper torso tensing or movements

4.9. repeated shoulder jerks

4.10. repeated tensing of abdominal muscles

4.11. wide-opening or rolling eyes repeatedly

4.12. repeated tensing or movements of hands or arms

4.13. any other motor tics

4.14. Other motor tics description

4.15. How long do/did the involuntary tics continue

4.15.a.1. For less than one month and gone now

4.15.a.2. For less than six months and still present

4.15.a.3. For less than one year and gone now

4.15.a.4. For more than a year and still present or gone now

5. Have you ever had involuntary tics, that started before the age of 18, and included repeated vocalization or sound-making?

6. If yes to question 5, do/did you have episodes

6.1. repeated grunting, whistling, or humming

6.2. repeated throat-clearing

6.3. repeated snorting or sniffing

6.4. repeating single words or syllables

6.5. any other vocal tics

6.6. Other vocal tics description

6.7. How long do/did the vocal tics continue

6.7.a.1. For less than one month and gone now

6.7.a.2. For less than six months and still present

6.7.a.3. For more than one year and gone now

6.7.a.4. For more than a year and still present or gone now

7. How have tics developed?

Appendix: Polycor correlation analysis:

To perform polycor correlation analysis, ordinal response data collected using TS/TD questionarie was subjected to hetrocorrelation coefficient estimates as an input for exploratory factor analysis (EFA). Therein, varimax rotation soultion was used to infer factors loading and structure. Factors with eigenvalues more than 1 were retained and characteristic consideration decided the final number of factors.

To eclucidate the validatity of predicted factors, confirmatory FA was employed using psych R package ‚polycor' (https://cran.r-project.org/package=polycor). Estimation was based on weighted least-squares and minimum residual calculation. Only items having factors loading greater than or equal to 0.4 were

retained in a factor (those with cross factors loading of greater than 0.3 were excluded from FA). Bayesian-information crierion, Tucker-Lewis Index were used as fitness indicess.

Supplementary Table 1: Demographic statistics of participant administered with TS/TD screening questionnaire.

Participant Group	*N*	Sex Males (Females)
ASD	266	152 (114)
ADHD	280	183 (97)
OCD	142	63 (79)
TS	191	104 (87)
TD	55	31 (24)
Relatives of ASD, ADHD, or OCD	3,286	1,722 (1,564)
General Population	211	109 (102)
Total	**4,431**	**2,364 (2,067)**

Supplementary Table 2: Scoring algorithm for the determination of TS and TD based on responses to screening questions (Appendix 1).

	Screening criteria	TD	TS
1	Any tic starting before the age of 18 and not due to other illness or medication	*	*
2	≥2 motor tics and ≥1 vocal tic with duration > 1 year	-	+
3	Motor tic(s) or vocal tic(s), not both >1 year	+	-
4	≥1 motor and/or ≥1 vocal tic only, duration > 4wks but less than a year	+	-
5	Self-report of diagnosed Tourette syndrome	-	+
6	Self-report of diagnosed Tic disorder	+	-

* "Yes" required for screening positive for TS or TD. TS is categorized based on criteria 1-4 or ‚yes' response to self reported clinical diagnosis in 5. TD positive case is registered where response ‚yes' to 1, and 3 or 4 or based on self reported clinical diagnosis in 6.

Supplementary Table 3: Tic disorders according to the DSM-IV-TR and ICD-10 diagnostic criteria.

Phenotype	DSM-IV-TR (code, label)	ICD-10 (code, label)	Diagnostic criteria*
Transient tic disorder	307.21, Transient tic disorder	F95.0, Transient tic disorder	Multiple motor and/or phonic tics with duration of at least 4 weeks, but less than 12 months. The tics usually take the form of eye-blinking, facial grimacing, or head-jerking.
Chronic motor or vocal tic disorder	307.22, Chronic tic disorder	F95.1, Chronic motor or vocal tic disorder	Either single or multiple motor or phonic tics, but not both, which are present for more than a year. The tics occur many times a day (usually in bouts) nearly every day or intermittently throughout a period of more than 1 year, and during this period there was never a tic-free period of more than 3 consecutive months.
Tourette syndrome	307.23, Tourette's disorder	F95.2, Combined vocal and multiple motor tic disorder (de la Tourette)	Both multiple motor and one or more vocal tics present, although not necessarily simultaneously. The tics occur many times a day (usually in bouts) nearly every day or intermittently throughout a period of more than 1 year, and during this period there was never a tic-free period of more than 3 consecutive months.
Tic disorder, unspecified	307.20, Tic disorder, not otherwise specified	F95.8, Other tic disorders and F95.9, Tic disorder, unspecified	Tics with short duration (i.e. less than 4 weeks); and onset of symptoms that occur after age 18 years

*To fulfull diagnostic criteria, onset of tics must be in childhood before the age of 18 (except F95.2, F95.8, F95.9) and tics should not be induced by medication or another medical condition.

Supplementary Table 4: Summary of exploratory factor analysis, using response data from TSQ data, showing factor loading by each tics category.

Tic Items	Body.Tics	Facial.Tics	Vocal.Tics
Leg.or.foot.movement	**0.73**	0.3	0.25
Abdomen.Tensing	**0.71**	0.32	0.2
Hand.movements	**0.7**	0.21	0.24
Shoulder.Jerks	**0.67**	0.44	0.03
Torso.Tensing	**0.65**	0.46	0.2
Clicking	**0.53**	0.33	0.24
Head.jerks	**0.52**	0.45	0.18
Facial.grimaces	0.43	**0.63**	0.09
Nose.twitch	0.4	**0.71**	0.21
Blinking	0.37	**0.76**	0.11
Eyerolling	0.36	**0.75**	0.19
Mouth.twitches	0.22	**0.8**	0.22
Wors.Syllables	0.3	0.13	**0.9**
Snore.Sniff	0.2	0.19	**0.85**
Grunt.Whistle.Hum	0.18	0.18	**0.81**
Throat.clearing	0.1	0.12	**0.87**

Supplementary Table 5: Summary of exploratory factor analysis showing variance explained and factor loading of different tics types.

Variable	Body Tics	Facial Tics	Vocal Tics
SS loadings	3.8	3.73	3.39
Proportion Var	0.24	0.23	0.21
Cumulative Var	0.24	0.47	0.68
Proportion Explained	0.35	0.34	0.31
Cumulative Proportion	0.35	0.69	1

MLE Chi Square = 392.44 with prob < 1.6e-44
Tucker Lewis Index of factoring reliability = 0.836
RMSEA index = 0.137 and the 90 % confidence intervals are 0.121 0.147
BIC = -17.35
Fit based upon off diagonal values = 1
Measures of factor score adequacy

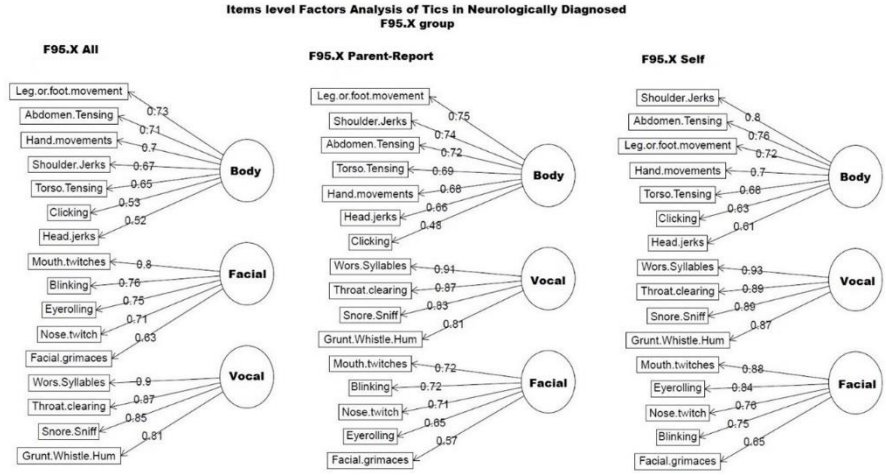
	Body Tics	Facial Tics	Vocal Tics
Correlation of scores with factors	0.90	0.92	0.97
Multiple R square of scores with factors	0.81	0.84	0.94
Minimum correlation of possible factor scores	0.62	0.68	0.88

Supplementary Table 6: Showing steps from CNV calling pipeline.

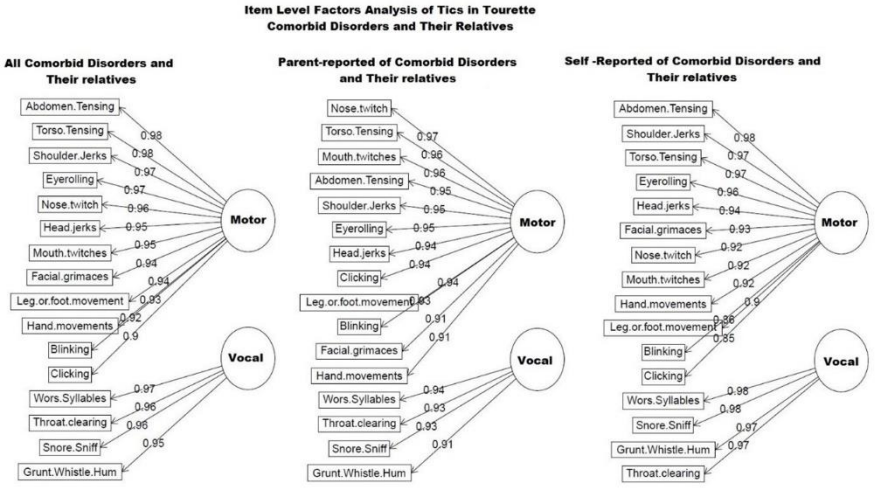
Step	Description	Total CNV calls	Unique CNV loci	Number of samples
CNV input	PennCNV predictions and validation through LRP data	6,973,363	1,306,432	150,656
Quality filtering	Removing CNVs overlapping gaps in the assembly	6,934,294	1,304,169	150,625
	Removing sample outliers (BAF-drift, LRR-SD and GCWF)	6,066,069	1,098,944	145,275
	Removing CNVs < 10 SNPs/Call	2,971,821	708,771	141,773
	Removing samples with number of calls > Mean+3SD	2,369,179	366,528	139,711
CNV validation	Autosomal CNVs segregating based on genealogy	1,300,158	142,534	136,601
	Segregating CNVs verified by LRP haplotypes	792,766	79,251	138,848
*CNV Breakpoint correction & Binning	Breakpoint correction & CNV Binning for LRP verified CNVs	24,282,133	87,464	138,848
CNV-bins with MAF > 0.01%	CNV with MAF > 0.01%	24,053,800	41,181	134,387

Supplementary Table7: Shows CNV stats for each chip.

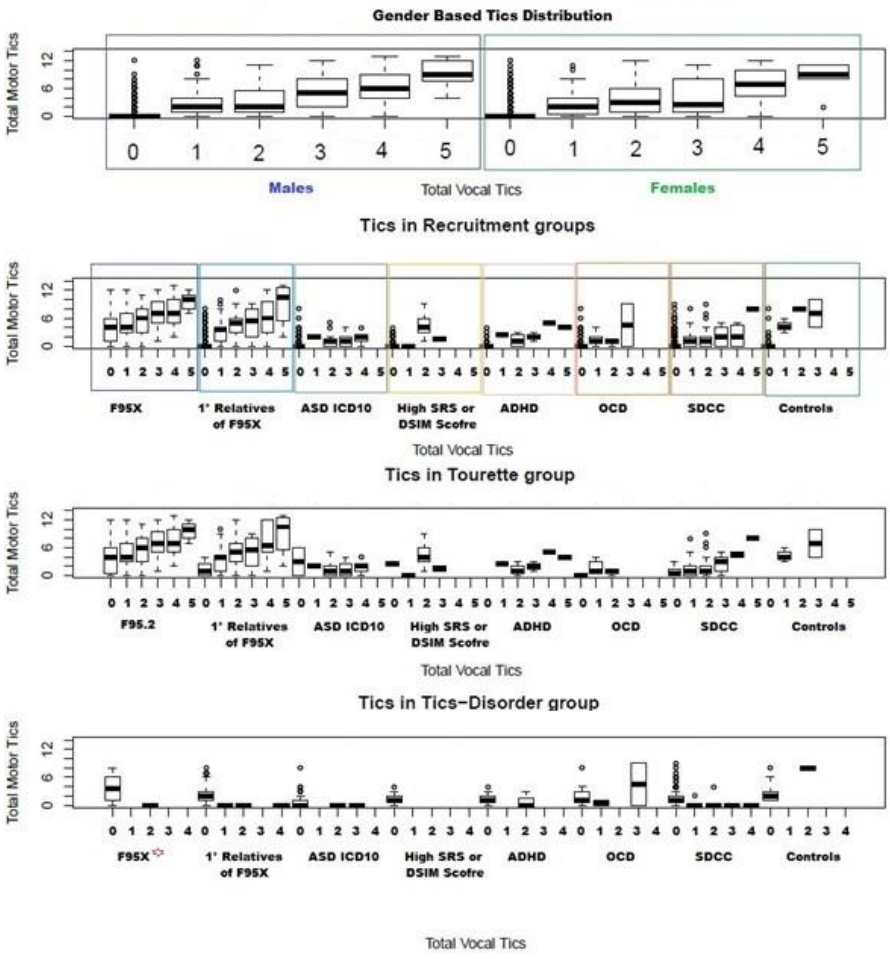
	Raw PennCNV		Calls after quality filtering	
Genotyping assay	CNV calls	samples	CNV calls	samples
DECODE_OEx-8_A	430,627	12,911	85,821	12,499
Human1M-Duov3_B	52,881	538	21,014	467
Human1Mv1_C	68,537	737	21,189	686
Human610-Quadv1_B	55,604	652	14,832	607
HumanCNV370-Quadv3_C	12,105	299	3,599	295
HumanCNV370v1_C	630,217	14,138	135,311	12,828
HumanHap300_(v1.0.0)	284,804	16,027	51,748	12,681
HumanHap300v2_A	87,483	6,744	13,518	5,187
HumanOmni1-Quad_v1-0_B	1,348,958	11,066	749,835	10,320
HumanOmni2.5-4v1_H	286,197	2,656	97,485	2,323
HumanOmni2.5-4v1-Multi_H	44,507	425	19,490	410
HumanOmni2.5-8v1_A	372,261	4,143	139,699	4,037
HumanOmni5-4v1_B	115,410	697	54,608	670
HumanOmniExpress-12v1_H	1,114,103	31,843	244,863	31,280
HumanOmniExpress-12v1-1_B	449,822	19,216	117,264	18,498
HumanOmniExpress-12v1-Multi_H	360,300	2,842	199,631	2,725
HumanOmniExpress-24v1-0_A	1,255,989	33,760	398,854	30,740
HumanOmniExpress-24v1-1_A	3,558	70	418	65
Total	**6,973,363**	**158,764**	**2,369,179**	**1,46,318**



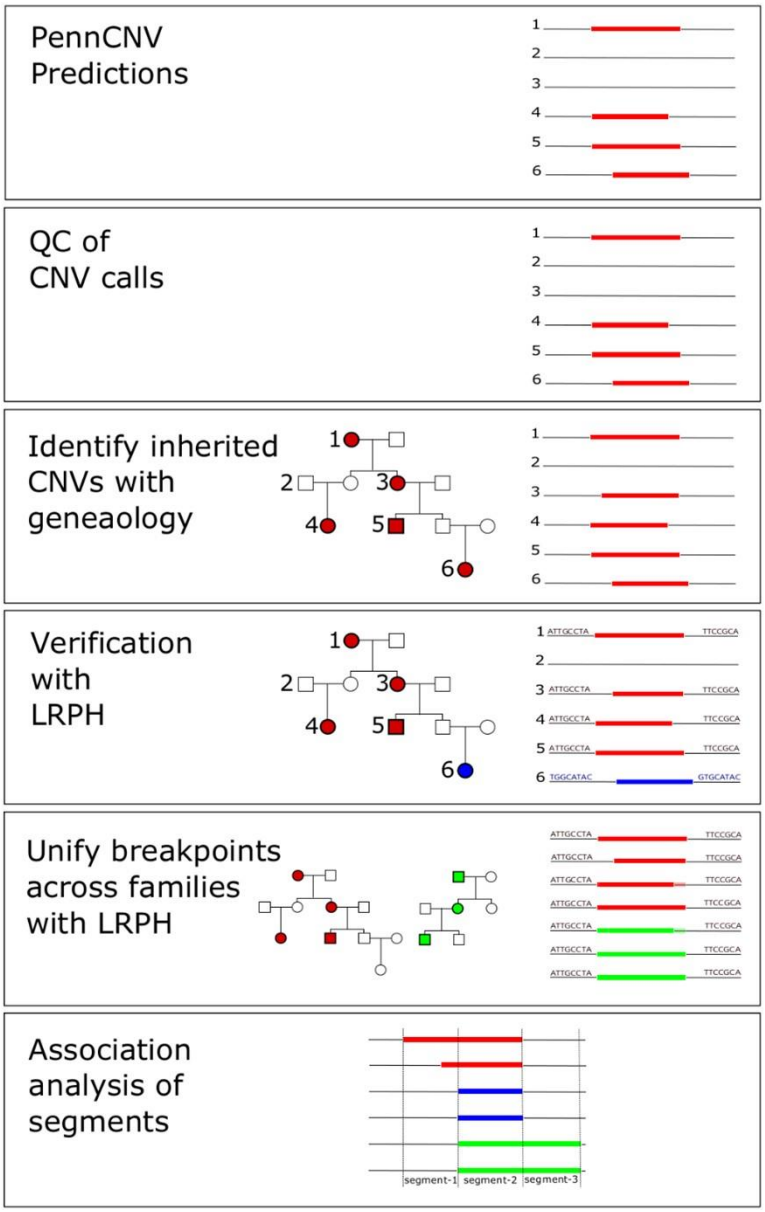
Supplementary Figure 1: Items level factors analysis of tics types in F95.* recruitment group. The tics response data was collected through TSQ.



Supplementary Figure 2: Items level factors analysis of tics types in recruitment groups excluding those with clinical diagnosis of F95.*. The tics response data was collected through TSQ



Supplementary Figure 3: Variable frquency of vocal and motor tics across genders and in different recruitment groups. SRS refers to social response scale, and SDCC is abbreviation for ‚State Diagnostic and Counselling Centre '.



Supplementary Figure 4: Flow chart of segregating CNV validation method, employing LRP approach, with exemplified CNV calls and pedigree information.