UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD Thesis

VIVEK APPADURAI

GENETIC ANALYSIS OF COMPLEX TRAITS IN POPULATION SCALE DATASETS

Supervisor: Professor Thomas Werge

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen December 15th, 2020

PREFACE

AUTHOR:

Vivek Appadurai | vivek.appadurai@regionh.dk

DEPARTMENT:

Institut for Biologisk Psykiatri, Psykiatrisk Center Sct. Hans, Boserupvej 2, Roskilde 4000

PhD SUPERVISOR:

Professor Thomas Werge | thomas.werge@regionh.dk

PRIMARY PhD CO-SUPERVISOR:

Dr. Alfonso Buil

CO-SUPERVISORS:

Dr. Wesley K. Thompson

Dr. Andrew J. Schork

FUNDING:

The Lundbeck Foundation

SUBMISSION DATE:

December 15 2020

ACKNOWLEDGEMENTS

First and foremost, I thank my mother, Mrs. Girija K. Rajagopal for the sacrifices she made throughout her life to provide a better education and quality of life for me than the one she received. Any work I have done so far and any future accomplishments in my scientific career or personal life will not be possible without her maternal, financial and moral support. I want to acknowledge my father, Mr. Subrahmanyam Appadurai for stressing the importance of ambition throughout my life. I'm grateful for my uncle, Mr. Sukumar K. Rajagopal for being an additional paternal influence in my life, especially in moments where I most needed his support.

I want to thank my supervisor, Professor Thomas Werge for funding my doctoral research and giving me the opportunity to work at the Institute of Biological Psychiatry (IBP), which opened doors and opportunities for learning and collaborative research I did not imagine were possible. I want to thank my co-supervisors Dr. Alfonso Buil and Dr. Wesley K. Thompson for recruiting me to IBP, identifying projects tailored to my skill set and providing timely feedback. I'm grateful for my co-supervisor, Dr. Andrew J. Schork who motivated me to pursue doctoral research and further acted as my day to day supervisor during my PhD. The mentorship and patience he afforded me during my steep learning trajectory in scientific research is immense and will have a lasting impact on my career. I wish to thank Mr. Patrick Minx and Dr. Wesley Warren at The Genome Institute at Washington University in St. Louis, who gave me my first opportunity in genomics and opened this career path.

I'm thankful for the friendship of Dr. Anders Rosengren and Dr. Dorte Helenius, with whom I shared an office all through my PhD. I'm grateful for the camaraderie of Dr. Ron Nudel, Dr. Sonja LaBianca, Xabier Calle, Dr. Morten Krebs and all my other lab mates at IBP. My research experience is richer for the ideas we shared and the motivation we offered each other. I'd like to especially thank the institute coordinators and administrative staff over the years at IBP including Kristian Krag, Søs Caspersen, Frants von Lüttichau and Marie Frost Arndal for handling my bureaucratic affairs and ensuring that my energy was conserved for research.

I express my gratitude to my co-authors, especially Dr. Lene Aaroe and Dr. Annette Erlangsen, for choosing to work with me and guiding me through the rigorous process of planning, analyzing and documenting scientific research. I thank Dr. Morana Vitezic and the bioinformatics team at Lundbeck A/S for hosting me during my change of research environment.

I finally thank the Lundbeck Foundation for making iPSYCH possible.

LIST OF PAPERS

PAPERS INCLUDED IN THIS THESIS:

1. Erlangsen A[a], **Appadurai V[a]**, Wang Y, Turecki G, Mors O, Werge T, Mortensen PB, Starnawska A, Børglum AD, Schork A, Nudel R, Bækvad-Hansen M, Bybjerg-Grauholm J, Hougaard DM, Thompson WK, Nordentoft M[b], Agerbo E[b]. ***Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study***. Mol Psychiatry. 2020 Oct;25(10):2410-2421. doi: 10.1038/s41380-018-0218-y. Epub 2018 Aug 16. PMID: 30116032; PMCID: PMC7515833.
2. Aarøe L[a], **Appadurai V[a]**, Hansen K.M, Schork AJ, Werge T, Mors O, Børglum AD, Hougaard DM, Nordentoft M, Mortensen PB, Thompson WK, Buil A, Agerbo E, Petersen MB. **Genetic predictors of educational attainment and intelligence test performance predict voter turnout**. Nat Hum Behav (2020). https://doi.org/10.1038/s41562-020-00952-2
3. **Appadurai V**, Rosengren A, Bybjerg-Grauholm J, Ingason A, Buil A, Mors O, Børglum AD, Hougaard DM, Nordentoft M, Mortensen PB, Werge T, Delaneau O, Schork AJ. **Legacy data, whole genome imputation and the analysis of complex traits: Lessons from the iPSYCH case-cohort study**. Manuscript in preparation.

[a] Shared first authorship

[b] Shared last authorship

PAPERS CO-AUTHORED DURING THE PhD:

1. Yapici-Eser H, **Appadurai V**, Eren CY, Yazici D, Chen C-Y, Öngür D, et al. Association between GLP-1 receptor gene polymorphisms with reward learning, anhedonia and depression diagnosis. Acta neuropsychiatrica. 2020:1-8.
2. Nudel R, **Appadurai V**, Schork AJ, Buil A, Bybjerg-Grauholm J, Børglum AD, et al. A large population-based investigation into the genetics of susceptibility to gastrointestinal infections and the link between gastrointestinal infections and mental illness. Human genetics. 2020:1-12.
3. Liu X, Nudel R, Thompson WK, **Appadurai V**, Schork AJ, Buil A, et al. Genetic factors underlying the bidirectional relationship between autoimmune and mental disorders–findings from a Danish population-based study. Brain, Behavior, and Immunity. 2020.
4. LaBianca S, Labianca J, Pagsberg AK, Jakobsen KD, **Appadurai V**, Buil A, et al. Copy Number Variants and Polygenic Risk Scores Predict Need of Care in Autism and/or ADHD Families. Journal of Autism and Developmental Disorders. 2020.
5. Schork AJ, Won H, **Appadurai V**, Nudel R, Gandal M, Delaneau O, et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. Nature neuroscience. 2019;22(3):353-61.
6. Nudel R, Wang Y, **Appadurai V**, Schork AJ, Buil A, Agerbo E, et al. A large-scale genomic investigation of susceptibility to infection and its association with mental disorders in the Danish population. Translational psychiatry. 2019;9(1):1-10.
7. Mullins N, Bigdeli TB, Børglum AD, Coleman JR, Demontis D, Mehta D, et al. GWAS of suicide attempt in psychiatric disorders and association with major depression polygenic risk scores. American journal of psychiatry. 2019;176(8):651-60.
8. Liu X, Helenius D, Skotte L, Beaumont RN, Wielscher M, Geller F, et al. Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration. Nature communications. 2019;10(1):1-13.
9. Mullins N, Bigdeli TB, Børglum AD, Coleman JR, Demontis D, Fanous AH, et al. Genome-wide association study of suicide attempt in psychiatric disorders identifies association with major depression polygenic risk scores. BIoRxiv. 2018:416008.

10. Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science. 2018;359(6376):693-7.
11. Fan CC, McGrath JJ, **Appadurai V**, Buil A, Gandal MJ, Schork AJ, et al. Spatial fine-mapping for gene-by-environment effects identifies risk hot spots for schizophrenia. Nature communications. 2018;9(1):1-7.
12. Benros ME, Nudel R, Wang Y, **Appadurai V**, Schork AJ, Agerbo E, et al. GWAS of the association between infections and mental disorders including heritability estimation and polygenic risk score analysis. Neurology, Psychiatry and Brain Research. 2018;29:5.

PRIOR PUBLICATIONS:

1. Besse A, Petersen A, Hunter J, **Appadurai V**, Lalani S, Bonnen P. Personalized medicine approach confirms a milder case of ABAT deficiency. Molecular brain. 2016;9(1):93.
2. Stiles AR, Ferdinandusse S, Besse A, **Appadurai V**, Leydiker KB, Cambray-Forker E, et al. Successful diagnosis of HIBCH deficiency from exome sequencing and positive retrospective analysis of newborn screening cards in two siblings presenting with Leigh's disease. Molecular genetics and metabolism. 2015;115(4):161-7.
3. Dogruluk T, Tsang YH, Espitia M, Chen F, Chen T, Chong Z, et al. Identification of variant-specific functions of PIK3CA by rapid phenotyping of rare mutations. Cancer research. 2015;75(24):5341-54.
4. **Appadurai V**, DeBarber A, Chiang P-W, Patel SB, Steiner RD, Tyler C, et al. Apparent underdiagnosis of cerebrotendinous xanthomatosis revealed by analysis of~ 60,000 human exomes. Molecular genetics and metabolism. 2015;116(4):298-304.

SUMMARY IN ENGLISH

Complex trait analysis methods have been employed to help elucidate the genetic architecture of polygenic traits like human behavior. Historically, these studies have been conducted through well-planned cohort designs and uniform data generation processes. The advent of genetic data generated through national biobanks, direct to consumer genetic testing companies, hospital systems and large research consortia has resulted in a multi-fold increase in sample sizes, which has provided a vital boost in the statistical power required for understanding the underlying biology of complex traits. Linking these genetic datasets to electronic health records and national patient registers has provided new research paradigms, where data ascertained for the study of particular outcomes can be brought to bear on a near infinite number of secondary outcomes. This data specific approach, in which valuable legacy data is repurposed for and extended to multiple aims, provides unique challenges and requires careful design. When considering genetic data resources, these legacy datasets often arise from multiple different technologies and their integration is critical to achieve large sample sizes. The complexity involved in integrating such legacy genetic data generated over multiple years and using diverse genotyping arrays raises novel challenges for current computational tools and protocols. The varied ascertainment of individuals presents unique challenges in the design of experiments to conduct complex trait analyses on secondary phenotypes.

This research thesis includes three manuscripts, all of which utilize the genetic data generated by the Lundbeck foundation initiative for Integrative Psychiatric Research (iPSYCH) Consortium.

The first paper in this thesis aims at understanding the genetic basis of suicide attempts, an unascertained, secondary phenotype in the iPSYCH cohort. It illustrates the importance of including diagnoses of ascertained severe mental disorders as covariates in complex trait analyses of secondary traits.

The second paper in this thesis investigates the predictive utility of genetic markers associated with educational attainment and intelligence test performance towards voter turnout in the 2013 Municipal, 2014 European and 2015 National elections in Denmark. We compare the extent to which this predictive utility is mirrored between a cohort of individuals with severe mental disorders, a population at risk of exclusion in the political process and a randomly ascertained nationally representative population of Denmark from the same time period.

The third paper in this thesis focuses on legacy genetic data integration directly. This work benchmarked the performance of existing computational tools and integration protocols for haplotype estimation and subsequent genotype imputation - two bedrock bioinformatics processes in the analysis of complex traits. It highlights the biases in complex trait analyses that can arise when integrating legacy data generated on two different genotyping arrays with minimal marker overlap. These problems appear exaggerated when current protocols are applied to non-European individuals in Denmark, highlighting an area for improvement before current research findings can equitably be applied in precision medicine initiatives.

DANSK RESUME

Der er anvendt komplekse analysemetoder til at belyse den genetiske baggrund for polygene træk og menneskelig adfærd. Historisk set er disse undersøgelser udført ved velplanlagte studie design og ensartede datagenereringsprocesser. Fremkomsten af genetiske data genereret gennem nationale biobanker direkte til forbrugernes genetiske testvirksomheder, hospitalssystemer og store forskningskonsortier har resulteret i en flerfoldig stigning i stikprøvestørrelser, hvilket har givet et afgørende løft i den statistiske styrke, der kræves for at forstå den underliggende biologi af komplekse træk. Sammenkædning af disse genetiske datasæt til nationale patientregistre giver også den enestående mulighed for at studere flere sekundære fænotyper bortset fra de træk af interesse. Kompleksiteten involveret i integrering af genetiske data genereret i store stikprøvestørrelser, der spænder over flere år og forskellige teknologier, skaber nye udfordringer for nuværende beregningsværktøjer og protokoller. Den varierede konstatering af individer præsenterer unikke udfordringer i designet af eksperimenter til at udføre komplekse trækanalyser på sekundære fænotyper.

Denne forskningsafhandling inkluderer tre manuskripter, som alle udnytter de genetiske data, der genereres af Lundbeck-fondenes initiativ til integrativ psykiatrisk forskning (iPSYCH) Consortium.

Den første artikel i denne afhandling sigter mod at forstå det genetiske grundlag for selvmordsadfærd, der kræver hospitalsindlæggelse, en ikke-kendt sekundær fænotype i iPSYCH-kohorten. Det illustrerer yderligere effekten af at inkludere diagnoser af konstaterede alvorlige psykiske lidelser som kovariater i estimater af SNP-arvelighed med snæver sans og genomvidere associeringsundersøgelser af ikke-selvmordsskader.

Den anden artikel i denne afhandling undersøger den forudsigelige nytteværdi af genetiske markører forbundet med uddannelsesmæssig opnåelse og intelligenspræstationspræstation i retning af valgdeltagelse i 2013 kommunale, 2014 europæiske og 2015 nationale valg i Danmark. Vi sammenligner i hvilket omfang denne

forudsigelige nytte konvergerer mellem en kohorte af individer med alvorlige psykiske lidelser, en befolkning med risiko for eksklusion i den politiske proces og en tilfældigt fastslået repræsentativ befolkning i Danmark fra samme tidsperiode.

Den tredje artikel i denne afhandling sigter mod at benchmarke præstationen af eksisterende beregningsværktøjer og fremhæve forspændingerne i komplekse egenskabsanalyser, der kan opstå, når man integrerer ældre data. Manglen på effektivitet af eksisterende protokoller og dataressourcer, når de anvendes til populationer af ikke-europæiske forfædre, fremhæves som et forbedringsområde, inden aktuelle forskningsresultater ligeligt kan anvendes i præcisionsmedicinske initiativer.

TABLE OF CONTENTS

1. MOTIVATION

The drop in costs of genotyping and next generation sequencing has led to a rapid proliferation of large genetic datasets, providing researchers with the kind of sample sizes needed to make impactful inferences from the analysis of complex traits. The cost of genetic analysis has moved from data generation to data storage, computational resources, and the bioinformatics expertise required to generate meaningful insights from such datasets[1]. The development of computationally efficient open source tools such as PLINK[2], GCTA[3], web interfaces like LDHub[4], FUMA[5] , GWAS Atlas[6], large imputation servers like the Michigan Imputation Server[7], along with user friendly statistical programming languages like R (https://www.r-project.org/) have made complex trait analysis more accessible to researchers and analysts with diverse expertise. However, there is a lack of research on the computational and bioinformatics practices for integrating legacy datasets generated over time and on different genotyping arrays and the biases such choices could introduce in the analyses of complex traits. The wealth of genetic and phenotypic data generated by biobanks and large genetic research consortia also provide cost effective ways to identify and analyze cohorts for multiple secondary traits other than the ones planned during the ascertainment of the study population. Choices in study design that do not account for ascertainment biases could lead to errors in estimates generated by such studies. If the estimates from complex trait analysis studies are to be used in precision medicine initiatives, it is imperative to characterize such biases prior to them being used in predictive modeling. In this thesis, I perform complex trait analysis using The Lundbeck foundation initiative for Integrative Psychiatric Research (iPSYCH) dataset[8], one of the largest cohorts designed for the study of severe mental disorders along with diagnostic and demographic information available from the Danish national psychiatric central[9] and civil registers[10].

In the first of three papers, I contribute towards the understanding of the genetic basis of suicide attempts, a secondary trait of interest in the iPSYCH cohort. In the second paper, I contribute towards demonstrating the predictive utility of genetic instruments associated with educational attainment and intelligence test performance towards voter turnout and the consistency of this predictive performance between one subset of iPSYCH, which was

ascertained as a nationally representative population of Denmark and the other, which was ascertained to include individuals with severe mental disorders. In the final paper, I contribute towards benchmarking tools for haplotype estimation and investigation of the best bioinformatics approaches to integrate legacy genetic data, along with the impact these choices could have on the analysis of complex traits.

2. RESEARCH OBJECTIVES

PAPER 1

- Estimate the variation in suicide attempts requiring hospitalization in the iPSYCH2012 cohort that can be attributed to variation in common single nucleotide polymorphisms.
- Perform genome wide association studies of the incidence of suicide attempts in the iPSYCH2012 cohort with and without adjusting for diagnosis of mental disorders as covariates.
- Perform a genome wide association study to identify genetic loci associated with comorbid occurrence of affective disorder and suicide attempts.
- Characterize the loci surpassing the threshold for genome wide significance.

PAPER 2

- Estimate the variance explained in voter turnout in the 2013 municipal elections, 2014 European elections and the 2015 national elections in Denmark by common single nucleotide polymorphisms in the nationally representative and psychiatric cohorts of iPSYCH2012.
- Estimate the predictive utility of genetic instruments associated with educational attainment and intelligence test performance towards voter turnout.
- Explore the genetic correlations of voter turnout with socio-economic, psychiatric, reproductive, cognitive and health related traits.

PAPER 3

- Benchmark the performance of haplotype estimation methods across four protocols for incorporating legacy data from individuals genotyped using two different arrays with minimal marker overlap.

- Estimate the biases that could arise from different choices of tools and data merging scenarios in the analysis of complex traits.

3. INTRODUCTION

3.1 COMPLEX TRAITS

Human traits can broadly be classified according to the underlying genetic architecture as monogenic, where most of the trait influencing variation is concentrated in a single gene, oligogenic, where a few genes confer strong effects towards the incidence of the trait with several others conferring comparatively weaker effects and polygenic, where the genetic loci influencing the trait are spread throughout the human genome and each locus contributes a small additive[11] effect to the overall genetic load[12].

In the initial years of genetic research, there was a strong debate between biologists studying discrete characters who favored mendelian principles of inheritance and biometricians who followed Galton's ideas in arguing that the genetic nature of continuous human traits like height could not be explained by mendelian inheritance patterns with single genes conferring large effects[13]. These two schools of thought were brought together when RA Fisher in 1918[14] proposed the infinitesimal model, suggesting that a very large number of mendelian-acting genes could explain the inheritance of biometrical traits under Mendel's principles of particulate inheritance. He further showed how this variation could be decomposed into variance arising from genetic and environmental factors.

S Wright in 1926[15], Dempster and Lener in 1949 [16] and DS Falconer in 1965[17] extended RA Fisher's infinitesimal model to dichotomous traits, proposing that the predisposition towards such a trait can be expressed as a combination of baseline genetic factors and a lifetime environmental exposure, termed the liability towards the trait. In a population, this liability is modeled as a normally distributed variable and the point above which the trait expresses in individuals is termed the threshold for the trait. The importance of genetic liability is shown as a shift in the mean of this distribution in relatives of individuals with a particular trait as compared to the general population. This is the liability-threshold model that links the theory of complex quantitative traits to complex discrete traits, such as common

disorders or diseases. The proportion of variance in a trait that can be accounted for by genetic factors was termed the coefficient of genetic determination or heritability[18].

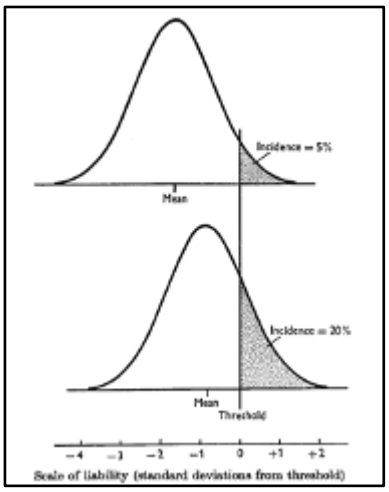


Figure 1. The normally distributed underlying liability for a trait in the general population (**top**) and in relatives of individuals with a particular trait (**bottom**). The vertical line indicates the threshold above which the incidence of the trait occurs and the spotted dots indicate individuals with the trait. ***Figure borrowed from Falconer DS, Ann. Hum. Genet., London. (1965), 29, 51[17]***.

Building upon the infinitesimal model, complex trait research in recent years has converged on the conclusion that the genetic architecture of complex traits is highly polygenic[19], meaning the trait enhancing loci are spread throughout the human genome, with each locus lending a small effect to an individual's baseline genetic load towards a trait or a disease. For example, Loh et. al 2015 inferred that at least 71% of 1mb regions in the human genome harbor a susceptibility variant for schizophrenia[20]. Due to the large number of trait susceptibility loci and the small effect conferred by each locus towards disease liability, combined with the fact that the manifestation of the trait itself depends on both genetic and environmental factors, these loci tend to be present in populations at relatively common allele frequencies despite selective pressure[21]. This is further called the common disease common variant hypothesis of complex traits[22].

Mental disorders such as schizophrenia[23], autism spectrum disorder[24], attention deficit hyperactivity disorder[25], depression[26] and socio economic traits such as educational attainment[27], intelligence test performance[28] and behavioral traits such as neuroticism, openness to experience, agreeableness, conscientiousness and extraversion[29] have all shown to be complex polygenic traits with varying degrees of genetic contribution. More recently, Kong et. al 2018[30] showed that the environmental contribution towards a trait can further be influenced by the genetics of the parents through a phenomenon they termed genetic nurture.

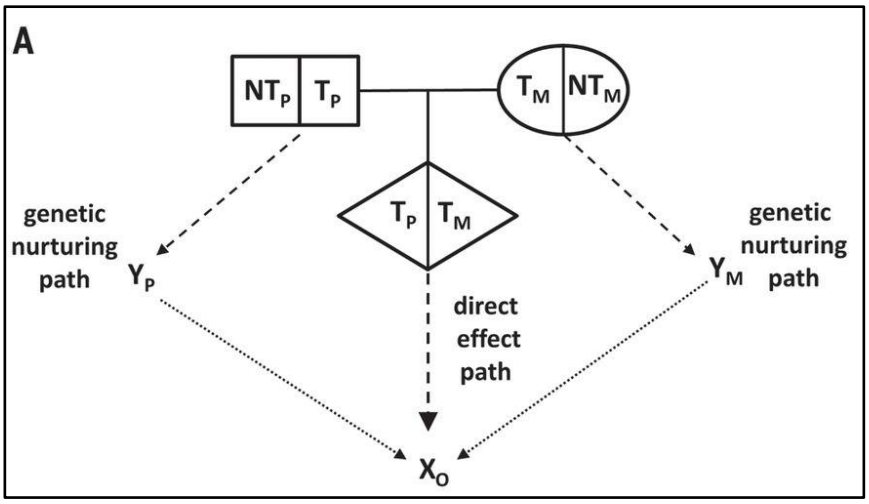


Figure 2. The direct genetic path influencing a trait x_0 in the offspring through the transmitted paternal and maternal alleles, T_P, T_M and the indirect genetic nurturing path acting through the paternal and maternal traits Y_P, Y_M, which are influenced by the non-transmitted paternal and maternal alleles, NT_P, NT_M. ***Figure borrowed from Kong et. al., Science 359, 424-428 (2018)[30]***.

Since the genetic susceptibility of an individual towards a trait or disorder remains constant throughout their lifetime, the knowledge of this baseline susceptibility attained through complex trait analysis methods can be used in genetically informed health care and policy making systems to provide timely interventions that modify environmental factors and improve quality of life[31]. The analysis of complex traits offers multiple benefits ranging from a better understanding of the biological pathways influencing a trait to identification of shared genetic risk factors between traits to finally predicting the incidence of a trait in an individual when genotype information is available.

3.2 METHODS USED FOR IDENTIFYING TRAIT ASSOCIATED LOCI

Initial attempts at uncovering genes associated with complex traits borrowed approaches from mendelian genetics with the usage of linkage analysis. In this approach, pedigrees with family members spanning multiple generations affected by a trait were genotyped and analyzed to identify segments of the genome co-segregating with the trait of interest. While these study designs proved moderately successful in identifying gene associations for traits such as Crohn's disease[32,33], Alzheimer's disease[34], type 2 diabetes[35,36], familial hyperlipidemia[37], when faced with truly polygenic traits like heart disease[38] and autism[39] the limitations of this approach in the face of locus heterogeneity and the limitations arising from the size of pedigrees required to afford enough power to detect associations became apparent.

Genome wide association studies (GWAS), built upon the common disease common variant hypothesis, provide a much finer scale map of trait associated loci. The GWAS era began in earnest[40,41] with the completion of the human genome project[42] which provided the necessary road map for organizing the database of single nucleotide polymorphisms (dbSNP)[43], a catalog of known genetic variation, and the HapMap project[44], a catalog detailing the linkage disequilibrium or the co-occurrence of alleles at two different sites in the human genome in different populations. These two catalogs, of known variation with predictable patterns of LD, allowed the design of genotyping arrays that provided efficient and expansive coverage of common genetic variations throughout the genome. The cost effective nature of these genotyping arrays has also made it possible to generate increasingly large genetic datasets at sample sizes required for the discovery of loci with small effects, such as those associated with complex traits.

In a genome wide association study, based on a case-control design, the outcome of interest is regressed against single nucleotide polymorphisms (SNPs) with socio-demographic factors such as age, gender and principal components of genetic ancestry as proxy for fine scale population structure as covariates to find SNPs that are present in higher frequency in a case group as compared to a control group. The effect sizes of each SNP towards the trait are the regression coefficients from this model and the statistical threshold for genome wide

significance is set at $p = 5x10^{-8}$ based on testing a million SNPs in accordance with LD patterns in the human genome observed in the hapmap project[45].

Thousands of variants have been identified for several common diseases using genome wide association studies and the associations have been catalogued in data repositories such as the GWAS catalog[46] and the GWAS Atlas[6]. The process of identifying the causal variant within the LD region of a locus that reaches statistical significance in association to a trait is termed fine mapping[47] and while fine mapping efforts have shown results for traits such as type 2 diabetes[48], this remains an area of active research. Assigning biological function to loci identified using genome wide association studies remains challenging, owing to much of the trait associated variation being present in non-coding regions of the genome[49], which are comparatively less well characterized[50]. However the summary statistics from genome wide association studies have proven immensely useful for estimation of heritability in absence of individual level genotype data, partitioning heritability in functional elements of the genome to search for enrichment[51], computing genetic correlations between various traits, calculation of risk profiles for patient stratification using polygenic scores and finally in inferring the mediation effect of one trait towards another using summary statistics based mendelian randomization. The concepts, strengths and limitations of genome wide association studies are further presented in detail in section 6.4.

Based on the design, genotyping arrays can give information on single nucleotide polymorphism (SNP) alleles in each individual at anywhere between 200,000 to 2 million loci in the human genome. In diploid organisms such as humans, each individual inherits a copy of a chromosome from each of their parents and each inherited chromosome can in turn be seen as a mosaic of several haplotype segments of correlated alleles in linkage disequilibrium (LD), that were inherited over several generations. Individuals of similar genetic origin tend to share more of these haplotype segments. The density and length of these shared haplotype segments become shorter with increase in population size[52]. Although the co-inheritance pattern of the alleles is lost in the genotyping process, this can be estimated using statistical methods, aided by a reference panel of haplotypes from a similar population. The statistical process of assigning each allele at a locus in an individual to a corresponding chromosome,

aided by haplotype information from large population reference panels is called haplotype estimation or phasing[53].

Since the loci genotyped on a genotyping array are unlikely to include all trait associated polymorphisms, genome wide association studies further rely on the LD between single nucleotide polymorphisms to find novel SNP associations with a particular trait. The single nucleotide polymorphism that is observed to be associated with a trait at a threshold passing the genome wide significance is quite often not the causal variant for that particular trait but is correlated with an unobserved causal variant due to both loci being inherited in the same haplotype block in linkage disequilibrium. Such SNPs are called tag SNPs[54,55]. As such, the chance of identifying trait associations increases with the increasing number of SNPs tested in an association study. Since most genotyping arrays lack the density to query enough of the regions in the human genome, a statistical technique called whole genome imputation[56] is employed to impute the SNPs at loci that were not genotyped by the array. This is done by using haplotype information estimated from phasing the SNPs that were genotyped on the array and a reference panel of dense haplotypes from an informative population. Whole genome imputation is presented in greater detail in section 6.3.

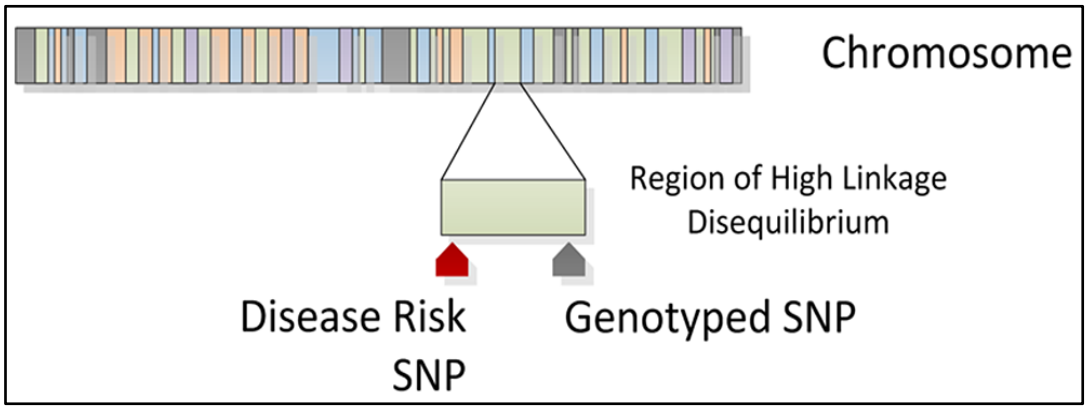


Figure 3. A genotyped SNP which acts as a tag SNP that is correlated with a disease risk SNP in a chromosomal region of high linkage disequilibrium. ***Figure borrowed from William S. Bush, Jason H. Moore (2012), Chapter 11: Genome Wide Association Studies PLOS Computational Biol 8 (12): e1002822[57].***

As the size of the haplotype reference panel and thereby the number of informative haplotypes increases, so does the accuracy of phasing and imputation[58]. Initially, the HapMap data with haplotypes from 1,301 individuals as of phase 3, was used as the haplotype reference panel for phasing and imputation and in subsequent generations, the 1000 genomes project[59], with haplotypes from 3,115 individuals as of phase 3, covering 26 different populations became the standard reference panel for genome wide association studies. More recently, the size of reference panels for phasing and imputation have increased owing to drops in costs of whole genome sequencing. Currently, the haplotype reference consortium (HRC) dataset with 64,876 haplotypes[60] and the Trans-Omics dataset for Precision Medicine (TOPMED)[61], consisting of haplotype information from approximately 155,000 individuals have become available for use either through imputation servers or for download through genetic data repositories. Imputation servers such as the Michigan Imputation Server have automated the process of phasing and imputation, thereby reducing the computational burden for complex trait research.

Association testing with copy number variants identified from genome wide SNP arrays as well as comparative genomic hybridization experiments have also revealed the contribution of rare variants to complex traits such as , schizophrenia[62,63], autism[64], anthropometric traits[65,66] and type 2 diabetes[67]. These are further cataloged in the database of genomic variants[68].

More recently, the reduction in costs of whole exome sequencing has also facilitated studies to examine the role of rare protein coding variation in complex traits such as autism, attention deficit hyperactivity disorder[69], schizophrenia and neurodevelopmental disorders[70]. As the allele frequencies in coding regions of the genome targeted by exome sequencing is usually rare, the association testing in such experiments involve tests of gene wise burden, comparing the allele count of protein altering variants in a gene, pathway or other functional element, ascertained by variant annotation between cases and controls[71].

3.3 SAMPLE SIZES FOR COMPLEX TRAIT ANALYSIS

While genome wide association studies with sample sizes in the tens of thousands identified several single nucleotide polymorphisms passing the threshold for genome wide significance, together these loci accounted for only a small proportion of variance even for highly polygenic traits. For example, the genome wide association study of height in approximately 30,000 individuals identified 27 significant loci which accounted for 3.7% of the variation in the trait[72], while twin studies suggested a genetic contribution exceeding 80%[73]. This led to discussion dubbed "the missing heritability problem[74,75], with speculation regarding the potential prevalence of dominance and epistatic interactions between genetic loci[76] and the extent to which rare variants that are not tested in genome wide association studies could explain the variation in complex traits. However, Yang et. al., in 2010[77] showed that when considering the additive effects of single nucleotide polymorphisms, of all variants, irrespective of the statistical significance of their association from genome wide association studies, the variance explained by single nucleotide polymorphisms, termed the SNP heritability, reaches 45% of the total trait variance. This implies complex traits were orders of magnitude more polygenic than had been previously thought and our expectations surrounding gene finding would need to be recalibrated. This is because as a trait is affected by greater numbers of variants, intuitively, their average effects are smaller and smaller. Concordantly, genome wide association studies at even larger sample sizes are needed to identify more of the trait associated loci at the stringent thresholds set for genome wide significance.

3.4 POPULATION SCALE DATASETS

Genetic datasets at large sample sizes with the requisite power to detect associations with complex traits can be accomplished through research consortia, where several smaller studies are meta-analyzed to enhance statistical power. More recently, several national, commercial, research and philanthropic initiatives aimed at uncovering the genetic basis of common diseases have led to sequencing of populations at large scale with cohort sizes reaching several hundreds of thousands of individuals. Detailed catalogs of phenotypic information on these individuals are further linked to the genetic data, either through

participant questionnaires, electronic health records from insurance companies, or, in the case of countries with single payer health systems, through national patient registers. These registers have the additional advantage of longitudinal information that may cover all the diagnoses an individual and their relatives have received over their lifetimes. While deCode genetics in Iceland[78] has been a pioneer in sequencing a significant share of the country's population, the UK Biobank[79], Finngen (https://www.finngen.fi/), Danish Blood Donor Study[80], the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) consortium[8], the million veteran program[81], Biobank Japan[82], BioVU[83] are more recent examples of such cohorts. Apart from these initiatives, direct to consumer testing companies like 23andme and AncestryDNA have also amassed cohorts with a large percentage of the consumers consenting to use of their genetic information towards research initiatives. Due to the magnitude of these datasets, the preferred and cost effective option for genetic data generation has been through genotyping arrays. As such, scientific perspectives built around these data sets and the intricacies of their assemblage and analysis are critical for advancing progress in modern human complex trait genetics.

3.5 OPPORTUNITIES FOR COMPLEX TRAIT ANALYSES

The emergence of large genetic datasets have resulted in a rich period of discovery for genetic variants associated with complex traits and diseases. The sample sizes of genome wide association studies in particular have risen steadily over the years owing to contributions from such cohorts. For example, the sample sizes for genome wide association studies of height undertaken by the Genetic Investigation of Anthropometric Traits consortium rose from 253,288 individuals in 2014[84] to 700,000 in 2018[85] with the introduction of genetic data from the UK biobank, thereby increasing the number of SNPs surpassing genome wide significance from 697 to 3,290. As an even starker example, the sample size for the genome wide association study of major depressive disorder leaped from 9,240 cases with no significant findings in 2012[86] to 135,458 cases in 2018[87] with the introduction of the UK biobank, iPSYCH and 23andme cohorts, thus revealing 44 independent loci surpassing genome wide significance. The wealth of phenotypic information aggregated by these datasets aid the design of studies that uncover the shared genetic origins of complex traits such as psychiatric disorders[88] and open up avenues for phenome-wide association studies

where multiple phenotypes can be investigated for association with a single genetic instrument[89].

3.6 CHALLENGES OF INTEGRATING LEGACY DATA

Owing to the recruitment and sample incorporation strategies of population scale research initiatives, genotyping is often carried out in multiple stages, spanning several years, often using the latest technologies in genotyping arrays. Due to differences in array design and manufacturing specifications, the number of overlapping markers between these different arrays might not be sufficiently large and hence, there arises an issue of missing information when the data is integrated. The conventional approach to integrating such datasets has been through haplotype estimation, followed by whole genome imputation, using a reference panel of ancestry informative markers from a large number of individuals who genetically resemble the study population so that all batches are imputed to the same marker set, followed by meta-analysis of GWAS summary statistics[90]. However it has been shown that the accuracy of haplotype estimation in particular relies on the sample size of the reference and target datasets[58] and for population scale datasets, the number of study haplotypes is often larger than any available reference panel. Therefore, it makes intuitive sense to pool together as many samples as possible for haplotype estimation to obtain the most accurate genotypes for downstream genetic analyses.

Combining samples genotyped on different arrays with sparse marker overlap can either reduce the overall marker set if only markers genotyped on all arrays are chosen or it can lead to missingness in the merged dataset if markers genotyped across all arrays are retained. While the bioinformatics tools used for haplotype estimation have been developed over several years with code optimization and mathematical approximations to scale well to current study sizes[91–95], the robustness and accuracy of these tools have not been tested in situations that can arise when merging samples that are not uniformly genotyped using the same marker set. There is a need to empirically evaluate and benchmark the phasing and imputation accuracy in such data integration scenarios so that researchers can make informed choices for bioinformatics workflows suited to their data generation processes.

Previous research on the effects of integrating samples genotyped on diverse arrays in complex trait analyses have investigated the possibilities of economically reusing controls across association studies[96,97] and found that using only a sparse set of SNPs common to all arrays reduced the imputation accuracy, whereas comparing SNPs genotyped on one array and imputed on the other leads to increased type I error rates in association studies, which is reduced but not completely eliminated when imposing stringent post imputation quality controls at the expense of lower coverage and thereby lesser power to detect novel associations. Exploring these biases and losses of accuracy across different data integration protocols employed in population scale complex trait analysis projects is essential to get replicable estimates from association studies as well as achieving good predictive accuracy when calculating polygenic scores.

3.7 ANALYSIS OF SECONDARY TRAITS

A rich amount of phenotypic information over the lifetime of a participant can be amassed from electronic health records, national patient registers, or participant questionnaires. When these phenotypes are combined with the genetic information generated by population scale datasets, it is possible to study multiple secondary traits at large sample sizes even if the population itself was not ascertained for such studies. While this is an economical way of repurposing data for less well studied traits, these opportunities also come with challenges of ascertainment differences if the subset of the biobank is not representative of a general population. If the secondary trait of study is correlated with the ascertained traits, the effects obtained from the analysis of secondary phenotypes can suffer from biases[98]. Further, the interpretation of associations from such studies becomes non-trivial as it is difficult to know if the association is driven by correlation with the secondary trait or the trait for which the study sample was ascertained[99]. As an example, it has been noted that the UK biobank study suffers from participation bias[100] and that the socio-demographic characteristics of the participants themselves are not representative of the general population of the United Kingdom[101]. However, the data still offers unrivaled sample sizes for the genetic study of traits such as loneliness[102] and insomnia[103], which have socio-demographic correlations in incidence.

Large genetic studies of suicide attempts (presented in this thesis), anxiety disorder[104] as well as cannabis use disorder[105] were performed in the iPSYCH cohort, where the case population is ascertained for individuals with severe psychiatric illness. While it is imperative to seize such opportunities, it is also important to take the sample ascertainment into account in the design and analysis plans for such studies as the prevalence of suicidal behaviour[106], anxiety[107] and cannabis use in individuals with severe mental disorders is higher than in the general population[108,109].

Richardson et. al (2007)[110] propose two different solutions for the analysis of secondary phenotypes. The first approach suggests using the traits for which the study sample was ascertained as covariates in the logistic regression model of a genome wide association study of a secondary trait and the second approach includes using the selection probability of samples included in the study as weights in the logistic regression model when analyzing a secondary trait of interest. An example of the suggested weighting scheme is to assign a weight of 1 for cases and a weight equal to the number of non-cases divided by the number of ascertained population controls for controls. Inverse probability weighted regression has further been supported as an unbiased alternative for such studies[111] without loss of power while Zaitlen et. al (2012)[112] demonstrated that conditioning of correlated clinical factors as covariates increases the power of case-control association studies.

Accounting for sample ascertainment issues in analysis of secondary phenotypes by either assigning weights to samples in association studies or through informed covariate adjustment for other diagnoses or in case-cohort designs, conducting the analysis in the population representative subsample of the study can generate estimates which are easier to interpret and can lead to better out of sample replication.

4. DATASETS USED

The following are the major genetic datasets used for the work presented in this thesis.

4.1 iPSYCH

The Integrative Psychiatric Research (iPSYCH) consortium funded by The Lundbeck Foundation was established to identify the underlying genetic and environmental factors influencing severe mental health disorders[8]. The genetic dataset generated as part of iPSYCH2012 was built as a case-cohort sample nested within 1,472,762 individuals born between the 1st of May 1981 and the 31st of December 2005 in Denmark, with a known mother, alive and residing in Denmark at the end of their first birth year. The iPSYCH2012 sample includes a total of 86,189 individuals of which 57,377 individuals were ascertained as cases with one or more severe mental disorders including schizophrenia, autism, attention-deficit/hyperactivity disorder and affective disorder. The cohort includes a total of 30,000 individuals, representative of the population of Denmark. Genotyping was carried out using DNA extracted from dried blood spots, obtained from the Danish National Biobank[113] at The Broad Institute, Boston, MA USA in 26 waves using the Infinium PsychChip Array v1.0 (Illumina, San Diego, CA USA). The iPSYCH2015 sample is an extension of iPSYCH2012 and nested within 1,717,316 individuals who were born in Denmark between the 1st of May 1981 and the 31st of December 2008, with a known mother, alive at the end of their first birth year and residing in Denmark. The genetic dataset generated as part of the iPSYCH2015 sample consists of 33,345 cases with severe mental disorders and a further 15,756 cohort individuals ascertained as a representative population of Denmark. Genotyping on these individuals was carried out using DNA extracted from dried blood spots obtained from the Danish National Biobank at Statens Serum Institut, Copenhagen DK using the Illumina Global Screening Array v2.0 (Illumina, San Diego, CA USA).

4.2 TRIOS

The trios dataset includes a total of 128 parent-child trios where the children were ascertained for a diagnosis of autism or attention-deficit/hyperactivity disorder with both parents born in Denmark on or after the 1st of May 1981. The samples were genotyped using both the Infinium PsychChip v1.0 and the Illumina Global Screening Array v2.0.

4.3 THE PERSONAL GENOMES PROJECT-UK

The Personal genomes project-UK (PGP-UK) is a genetic data resource created to facilitate open access to multi-omics datasets towards gaining insights into biological and medical processes[114]. PGP-UK contains multi-omic data pertaining to a total of 1,100 individuals, who were either citizens or permanent residents of the United Kingdom, consented to sharing their biological information and further passed an entrance test aimed at knowing the risks of sharing genetic data. DNA was extracted from blood spots and whole genome bisulfite sequencing was performed using the Illumina HiSeq X Ten platform at an average depth of 15x. The resulting FASTQ and BAM files were deposited to the European Nucleotide Archive (ENA) with the study identifier PRJEB17529 (https://www.ebi.ac.uk/ena/browser/view/PRJEB17529).

4.4 THE DANISH CIVIL REGISTRATION SYSTEM

The Danish civil registration system was established in 1968[10] and through a unique 10 digit personal identifier, enables linkage of individual level information across the several Danish registers. This enables a highly accurate, anonymized way of ascertaining cohorts representative of the general population as well as for case-control studies, making the civil registers a powerful tool for epidemiological research conducted in the Danish population[115].

4.5 THE DANISH PSYCHIATRIC CENTRAL REGISTERS

The collection of clinical data on in-patient treatment in psychiatric hospitals in Denmark began in 1938 and this was curated into an electronic database, the nationwide

psychiatric central register in 1969[9]. The register contains records of every psychiatric treatment requiring admission to a hospital since 1969. The diagnosis records for severe psychiatric illnesses are indicated as a binary variable but no further information on the severity of the diagnosis is available but when used in conjunction with the Danish civil registration system, it is possible to ascertain the age at which an individual was diagnosed with a psychiatric illness. In 2000, the psychiatric central register was further split into the psychiatric central research register (PCRR) and the psychiatric central treatment register (PCTR) with the PCRR providing data for epidemiological research.

While the PCRR is comprehensive for all severe psychiatric illnesses such as schizophrenia, mild to moderate cases treated by general practitioners are not recorded in the PCRR. It is also possible for individuals with anxiety, affective and personality disorders to be treated by private psychiatrists and information on these diagnoses are unavailable in the PCRR.

5. ETHICAL CONSIDERATIONS

All research pertaining to iPSYCH has been conducted with prior approvals from the Danish Scientific Ethics Committee, Danish Health Authority, Danish Neonatal Screening Biobank Screening Committee. All individual level information was stored and analyzed on a secure server within the Danish National Life Science Supercomputing Center (Computerome) (https://computerome.dtu.dk/). The PGP-UK has been approved by the University College London Scientific Ethics Committee and conducted under the auspices of the national research laws and the Declaration of Helsinki.

6. METHODS

6.1 HERITABILITY

The heritability of a trait can be defined as the amount of phenotypic variation that can be explained by genetic factors. If V_P represents the net variation in a phenotype, it can be expressed as the sum of the variation in the individuals' genetics (V_G) and the variation in the environment (V_E).

$$V_P = V_G + V_E$$

Then the broad sense heritability of the trait, explained by all genetic factors can be expressed as:

$$H^2 = \frac{V_G}{V_P}$$

$$\Rightarrow H^2 = \frac{V_G}{V_G + V_E}$$

The broad sense heritability can further be decomposed into the variance that can be accounted for by additive genetic factors (V_A) and non-additive genetic factors (V_{NA}) such as dominance interaction within genetic loci or epistatic interactions among them.

$$V_G = V_A + V_{NA}$$

Therefore the narrow sense heritability can be expressed as:

$$h^2 = \frac{V_A}{V_A + V_{NA} + V_E}$$

It is important to note that the heritability of a trait is a property of a particular population at a particular time period and it cannot be used to stratify different populations

or across time[114]. The traditional methods for estimation of heritability have relied on twin datasets, where heritability can be estimated as the difference in phenotypic correlations between monozygotic twins, who share 100% of the genetics and environment and dizygotic twins who share 50% of the genetics and 100% of the environment. Other methods for estimating heritability include regressing offspring phenotype values against mid-parent phenotype values or with the use of linear mixed models in case of the availability of phenotypic information from relatives from extended pedigrees spanning multiple years. However, these methods can lead to biased estimates of heritability if the trait similarity between relatives is biased due to a shared environment[18].

The heritability that can be estimated from population based cohorts of unrelated individuals is termed narrow-sense SNP heritability (h^2_{SNP}), which is the amount of phenotypic variance that can be accounted for by single nucleotide polymorphisms that can be ascertained from genotyping arrays. As these arrays do not directly measure causal SNPs, it is further a fraction of the narrow-sense heritability (h^2). When the genetic information on a population of unrelated individuals is available, the pairwise relatedness between all individuals is estimated directly as similarity across the measured SNPs and stored in a genetic relatedness matrix (GRM). A linear mixed model using this GRM in place of the traditional kinship matrix is then employed to estimate the phenotypic variance explained by measured genetic relatedness among individuals. This method has been implemented in a tool for Genome Wide Complex Trait Analysis (GCTA)[3]. GCTA further allows the estimation of heritability on a liability scale when given the population prevalence of the trait. The work submitted as part of this thesis has extensively used GCTA for estimating narrow sense SNP heritability.

In the absence of individual level genetic information, the heritability can be estimated from summary statistics obtained from genome wide association studies using Linkage Disequilibrium Score Regression[116]. Linkage Disequilibrium Score Regression works on the assumption that the more SNPs a given SNP is correlated with, the higher the chance of it tagging a potential causal variant and such SNPs will thus have an elevated test statistic in genome wide association studies, and the elevation is expected to rise as a function of the effective number of SNPs it is correlated with. At each SNP, the test statistic from the GWAS

is regressed against the linkage disequilibrium score of the SNP and the slope of the regression line is the heritability explained per each SNP. The linkage disequilibrium score which is the sum of all correlations of the SNP with its neighboring variants is ascertained from a population of individuals genetically similar to the population used in the GWAS.

6.2 HAPLOTYPE ESTIMATION

Diploid organisms like humans receive a copy of each autosomal chromosome from each parent. The data from genotyping arrays and sequencing experiments reveal the allelic information at each locus, but the co-inheritance pattern of alleles across the different loci is lost. Haplotype estimation (phasing) is the process of linking alleles across diploid loci to chromosomes of origin[43]. In small populations with a large fraction of genotyped individuals, methods that leverage long identical by descent (IBD) segments shared between closely related individuals can be used to estimate haplotypes[117] and this approach can be highly accurate when one or more individuals in the population is homozygous in a genomic region. Phasing in unrelated individuals can be computationally achieved by leveraging cryptic relatedness that might exist even in seemingly unrelated individuals and with the aid of a reference panel of informative haplotypes, subsets of which can inform the haplotype segments of a target individual[58].

The three most prominently used tools for haplotype estimation in unrelated individuals from biobank scale datasets are EAGLE2[91], BEAGLE5[93] and SHAPEIT4[94]. All three tools use hidden markov model (HMM) approaches based on the Li and Stephens model[118], which gives a computationally scalable, probabilistic way of estimating a haplotype h_k as a mosaic of previously observed haplotypes h_1, h_2, h_3 .. h_{k-1} while accounting for genetic phenomena like mutation and recombination rates.

Figure 4. A simple representation of haplotype estimation using the Li and Stephens model. The row G indicates the allelic dosage (0, 1, 2) at a locus with the row D_1 indicating the two haplotypes colored in orange and green as a mosaic of H_k haps that have been observed. ***Figure borrowed from Delaneau et. al Nature Communications 10: Article Number: 5436 (2019)[94].***

The methods differ in the data structures they use to store the most informative haplotypes during haplotype estimation and the approaches they use for achieving memory run time efficiency and when presented with target and reference datasets with several hundreds of thousands of haplotypes. All three tools give the user the option to tune parameters for a choice between run time, memory usage and accuracy depending on the available computational resources. SHAPEIT4 uses the positional burrows wheeler transform[119] to index and rapidly retrieve a set of most informative haplotypes for every 2 megabase region in the genome. BEAGLE5 constructs a set of composite reference haplotypes at each phasing interval from the full set of haplotypes present in the reference while prioritizing haplotypes with longer identity by state segments. EAGLE2 uses the positional burrows wheeler transform to build HapHedge data structures for each chromosome where reference haplotypes carrying the same prefix are condensed down to a single branch with the branch carrying the frequency information for such prefixes. For each target sample, the most likely haplotype segment is selected by performing a beam search from the left to right of a HapHedge data structure. While the performance of these phasing tools have been tested for accuracy and scalability with reference and study panels of different sizes, the tests have

usually been performed under ideal conditions in well studied cohorts like subsets of the UK biobank, the genome in a bottle dataset[120] or a subset of the 1000 genomes dataset.

In the first and second papers presented in this thesis, we use SHAPEIT3[95] to perform haplotype estimation and the 1000 genomes phase 3 dataset[59] as the set of reference haplotypes. In the third paper presented in this thesis, titled "Legacy Data, Whole Genome Imputation and The Analysis of Complex Traits: Lessons from the iPSYCH Case-Cohort Study", we test the robustness of these tools under different data combination scenarios with different densities of input marker sets, target sample sizes and varying degrees of missingness as might be seen in large genetic research datasets with samples collected in batches over multiple years and using different genotyping arrays. We use the haplotype reference consortium (HRC) version 1.1[60] as the set of reference haplotypes.

6.3 WHOLE GENOME IMPUTATION

Genotyping arrays are designed to provide allelic information at anywhere between 200,000 to 2 million loci in the human genome. Missing data imputation, which follows haplotype estimation is the process of estimating alleles at non-genotyped loci in an individual using the sparse allelic information from loci that were genotyped and a reference panel of haplotypes[47]. Imputation, much like haplotype estimation relies on the basis that even unrelated individuals share segments of DNA in linkage disequilibrium, inherited from common ancestors.

Whole genome imputation is an essential step in complex trait analysis pipelines for increasing the number of markers to be tested and thereby increasing the power of genome wide association studies. Imputation also aids in fine mapping efforts to find the causal locus in a linkage disequilibrium block of SNPs associated with a particular trait, as well as in providing a common SNP set for meta-analysis of cohorts generated by different research groups and finally to enhance the marker overlap between the reference GWAS and target genotypes for calculation of polygenic scores[121].

The modern computational tools used for whole genome imputation[93,122] use hidden markov model (HMM) methods based on the Li and Stephens model, previously described in section 6.2 on haplotype estimation and a reference panel of dense haplotypes obtained from sequencing experiments to generate probabilistic estimates of alleles at missing loci in a target sample.

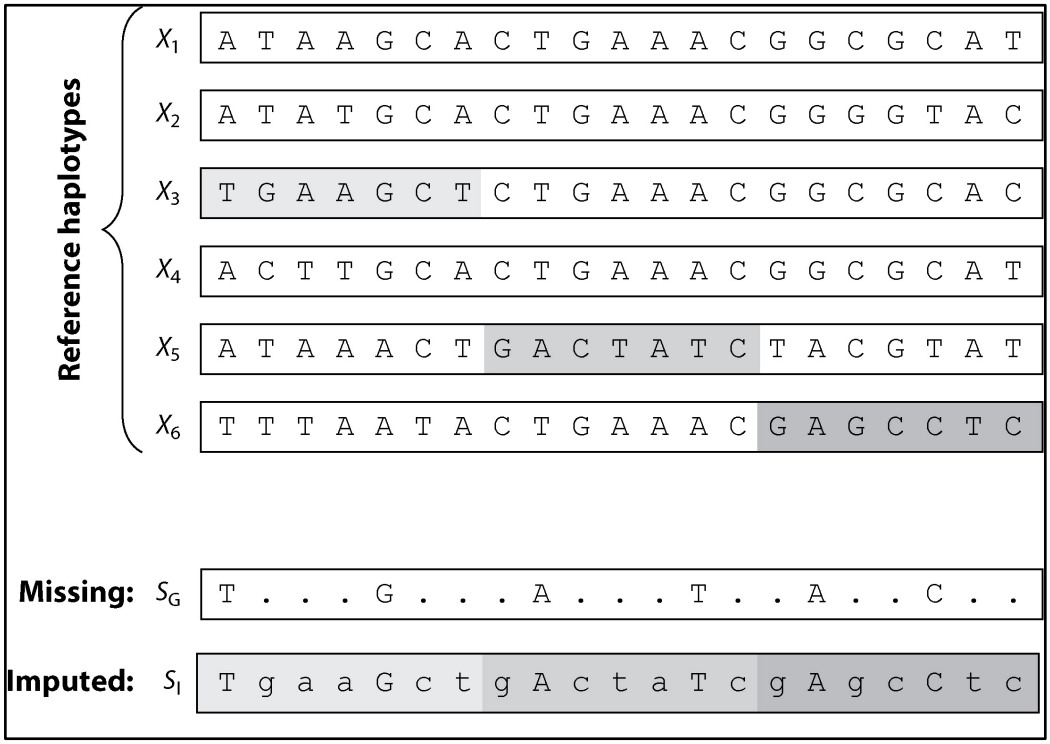


Figure 5. An imputed haplotype S_I is generated using information from a set of reference haplotypes, X_1 .. X_6 and a sparsely genotyped study haplotype, S_G with missing alleles. ***Figure borrowed from Das et. al Annu Rev Genomics Human Genet. 2018 Aug 31;19:73-96[121]***.

Imputation tools emit the certainty of the estimated allelic probabilities through metrics such as INFO scores or r^2, which is an estimate of the correlation between the true alleles and the posterior allelic probabilities at each locus estimated by imputation. It is common to filter imputed loci using a threshold for INFO scores or r^2. The r^2 metric can be interpreted as a reduction in effective sample size when testing trait association at a particular imputed locus. For example, if the r^2 at a locus is $\hat{r} \in [0, 1]$and the total sample size is N, the effective sample size at that locus is reduced to $\hat{r} * N$[123].

As haplotype estimation and whole genome imputation are both computationally intensive processes, research groups can avail themselves of web servers such as the

Michigan imputation server[7] to reduce the bioinformatics burden of complex trait analysis. However, the data protection stipulations governing storage and export of genetic data obtained through national biobanks might make the use of such a service prohibitive.

In the first two papers presented as part of this thesis, the imputation of missing genotypes was performed using Impute2[124] and using the 1000 genomes phase 3 reference panel[59]. In the third paper presented as part of this thesis, imputation of missing genotypes was performed using BEAGLE5.1[93]. Since the large population reference panels are predominantly based on haplotypes from individuals of European ancestry, the work presented as part of this thesis also evaluates the performance of missing data imputation in individuals of non-European ancestry in iPSYCH.

6.4 GENOME WIDE ASSOCIATION STUDIES

In protocols for genome wide association studies (GWAS), an initial set of quality control steps are undertaken to censor SNPs and individuals of poor quality due to technical artifacts in the data generation process. These quality control methods involve exclusion of SNPs that are found missing in > 5% of the individuals in the study population or displaying differential missingness between cases and controls. SNPs violating the Hardy Weinberg equilibrium in the control subset of the cohort are excluded. If genotyping of study subjects is carried out in batches, any SNPs showing association to a genotyping batch are excluded. SNPs at minor allele frequencies less than 1% are excluded from association testing as the accuracy of genotyping arrays is unreliable at low minor allele frequencies. For individual level quality control, study subjects missing more than 5% of the genotyped SNPs are excluded. Individuals showing abnormal levels of heterozygosity or runs of homozygosity that cannot adequately be explained by admixture are excluded due to potential sample contamination. The study sample set is further reduced to a subset of individuals of a relatively homogenous genetic origin using principal component analysis so that population stratification does not lead to false positives in SNP associations. The individuals are further pruned to exclude any relatives up to a 2nd degree of relatedness and GWAS is conducted on this set of unrelated individuals of a homogenous genetic origin.

Since GWAS relies on linkage disequilibrium between SNPs to identify associations with tag SNPs correlated with a potential causal SNP, the odds of finding trait associated loci increases when a dense set of markers are used for association testing. This is usually accomplished by phasing the few hundred thousand SNPs genotyped on an array and further imputing them to millions of SNPs using reference panels of dense haplotypes as described in sections 6.2, 6.3. Post imputation quality control includes excluding SNPs based on some threshold for imputation quality score, SNPs associating with an imputation batch, SNPs that violate Hardy Weinberg equilibrium in controls and SNPs with minor allele frequencies less than 1% as imputation quality degrades at rare minor allele frequencies.

The traditional model employed by genome wide association studies (GWAS) to test the association of each single nucleotide polymorphism (SNP) with a dichotomous trait of interest is as follows:

$$logit(p_i) = \beta_j X_{ij} + E$$

Where p_i = Probability that individual i has the trait.

$X_{ij} \in [0, 2]$ is the additive genotype dosage of individual i at SNP j

β_j= Effect size (natural log of odds ratio) of the SNP j towards the trait

E = the residual error term that captures the non-additive genetic effects

The significance of the effect is obtained by testing the alternative hypothesis $H_A: \beta_j \neq 0$against the null hypothesis $H_0: \beta_j = 0$

Sociodemographic variables such as age and gender, the first ten principal components of genetic ancestry to correct for population stratification[125] are often used as covariates in the regression.

The results of genome wide association studies are visualized using Manhattan plots with a horizontal line indicating the threshold for genome wide significance, set at $p = 5x10^{-8}$. Inflation in test statistics is calculated using genomic control which compares the median of the test statistics to a chi square distribution with one degree of freedom. This inflation is

visualized using quantile-quantile plots. While inflation in test statistics arising from population stratification or cryptic relatedness in the study cohort needs to be corrected, a certain degree of inflation in GWAS test statistics is expected for highly polygenic traits[23].

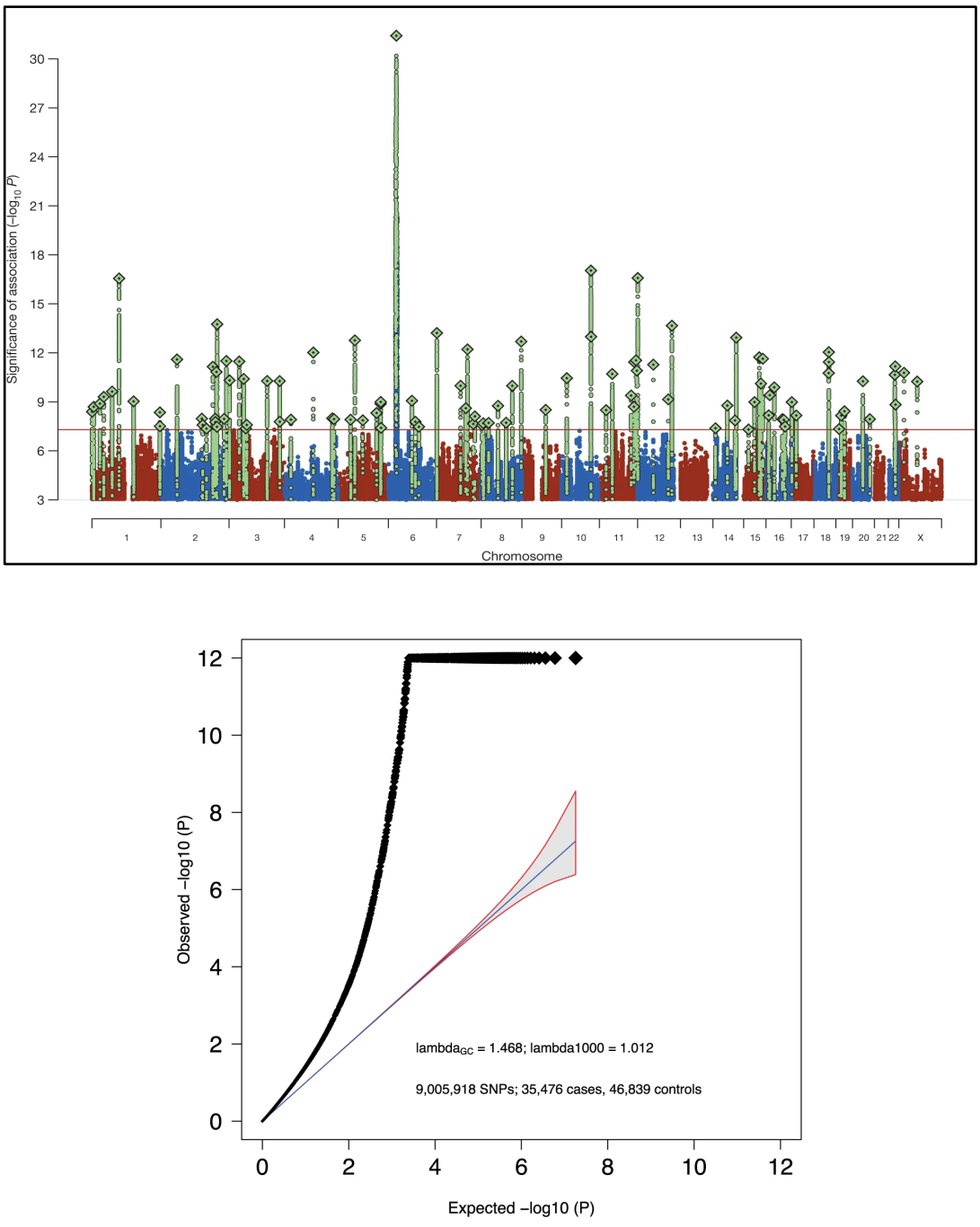


Figure 6. The Manhattan and quantile-quantile plots from the genome wide association study of Schizophrenia. ***Figure borrowed from Ripke S, Neale B, Corvin A et. al. Biological Insights from 108 schizophrenia associated genetic loci. Nature 511, 421-427 (2014)[23]***.

The quality control measures used for selecting high quality genotypes and study samples for use in genome wide association studies and other complex trait analyses conducted in this thesis were performed using PLINK[2]. Principal components were calculated using EIGENSTRAT[125] and FlashPCA[126]. Genome wide association studies were carried out using PLINK. Relatedness pruning was performed using KING[127].

6.5 POLYGENIC SCORES

Predicting an individual's likelihood of acquiring a trait when given their genotypes is of keen interest for not only complex trait analysis, but also precision health initiatives. This is currently accomplished through the calculation of polygenic scores, where the net score of an individual in a cohort towards a trait can be computed as the running sum of the product of the SNP effects obtained from an external genome wide association study and the additive genotype dosages for the individual at each SNP.

$$PGS_j = \sum_{i=1}^{m} \beta_i X_{ij}$$

Where PGS_j = Polygenic score of individual j

β_i= Effect size of SNP i towards the trait

$X_{ij} \in [0, 2]$ is the additive genotype dosage of the effect allele for individual j at SNP i

Khera et. al in 2018[128] calculated polygenic scores for five common diseases including coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer respectively and found that for coronary artery disease, the disease risk in individuals identified in the extreme tails of the PGS distribution was 20 fold higher than the risk for those identified as carriers of disease implicated rare monogenic mutations. They argue for the implementation of routine PGS screening in clinical practice so as to modify lifestyle factors and protect against disease incidence.

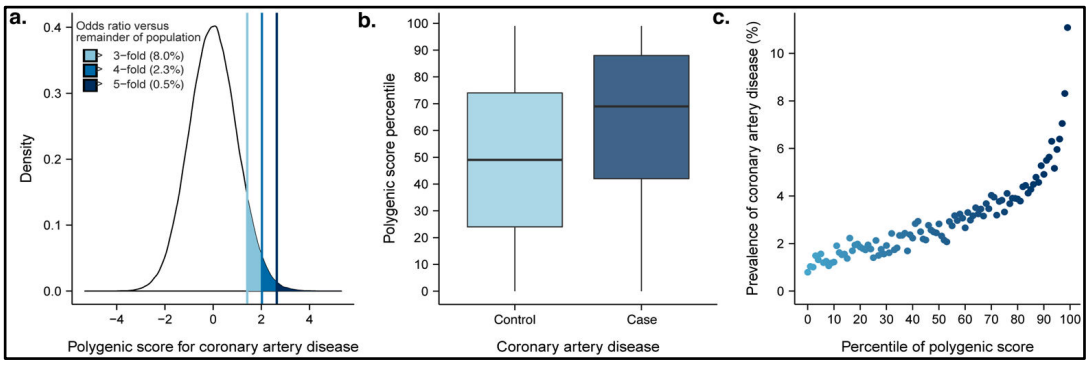


Figure 7. Panel (**a**) shows the distribution of polygenic scores for coronary artery disease with odds ratios of risk for individuals in each shaded region above a certain threshold score as compared to the rest of the cohort. Panel (**b**) shows the higher polygenic score for cases as compared to controls. Panel (**c**) shows the increment in prevalence of coronary artery disease with each percentile increase in polygenic score. ***Figure borrowed from Khera et. al., Nature Genetics 50, 1219-1224 (2018)[128].***

Computation of polygenic scores is an active area of research in the complex trait analysis community and there are several methods to compute scores varying by the choice of SNPs and the weight assigned to each SNP. Pruning and thresholding methods[129] exclude SNPs from the target dataset based on estimated linkage disequilibrium to choose a set of independent SNPs, weighed by effect sizes obtained from genome wide association studies. Other methods prune SNPs for independence using external linkage disequilibrium reference panels and assign posterior effect sizes estimated through bayesian frameworks, assuming that SNP effect sizes are drawn from either a normal distribution or from a mixture of distributions modeling the genetic architecture of the trait being predicted[130,131]. Functional annotations have also been used in cases where they can identify true positive associations in GWAS and assign a higher weight to such SNPs[132]. Methods that are used to evaluate the predictive performance of PGS include estimation of the variance in a trait explained by the PGS on a liability scale, calculation of area under the receiver operating curve, which are informative for how well the PGS is calibrated across different training and test datasets. Calculation of odds ratios for individuals in the top percentiles of a polygenic score distribution as compared to the rest of the population has utility in terms of prioritizing patients in high deciles of risk for clinical intervention and treatment.

Ni et. al in 2020[133] performed a comprehensive evaluation of PGS methods in psychiatric phenotypes and suggested that attenuation in the predictive accuracy of PGS can arise due to differences in the phenotype definition between the discovery GWAS and the target cohort, technical artifacts such as genotyping and imputation errors in the data generation process, differences in ethnicity between the reference GWAS and the target cohort[134] and variation in the ascertainment criteria of cases and controls between the reference GWAS and the target cohort. Since a vast majority of the genome wide association studies in current research are performed in individuals of European ethnicity, the predictive power of PGS for non-European populations lags in comparison and this could widen the existing healthcare disparities for minority individuals if PGS are to be incorporated in precision medicine initiatives[135]. Genome wide association studies being conducted in diverse data resources such as the Biobank of Japan are a step forward in bridging this gap. However more research is needed towards understanding the heterogeneity introduced by technical artifacts in the genotyping process and phenotype ascertainment in genome wide association studies towards polygenic scores to enhance their predictive utility across different datasets.

The work included in this thesis employs a pruning and thresholding method as implemented in PRSice[136]. First, the study genotypes are harmonized with the summary statistics obtained from the external GWAS, prior to linkage disequilibrium pruning to ensure a set of independent SNPs. We used all pruned SNPs overlapping the reference GWAS and the study markers to calculate polygenic scores. The predictive performance of the polygenic scores was analyzed by calculating the variance explained in the phenotype on a liability scale, estimated using trait prevalence in the randomly ascertained cohort of iPSYCH2012.

6.6 GENETIC CORRELATIONS AND PLEIOTROPY

Considering the polygenic nature of complex traits and that the trait associated loci are spread throughout the genome, it is conceivable that the same loci might influence more than one trait in a phenomenon called genetic pleiotropy[137]. Based on the mechanism of action of the pleiotropic loci, the pleiotropy can be horizontal, where the genetic locus independently influences two traits, vertical, where the influence of the locus on one trait

acts through a second trait, or it can be spurious where a genetic locus associates with both traits due to being in linkage disequilibrium with two different causal variants in a large LD block in the genome[138]. A genetic correlation is a quantitative measure of the shared genetic origins between two traits owing to the pleiotropic actions of genes. Suppose there are traits X and Y, each defined as the sum of genetic values (g_x, g_y) and residuals in the trait value not accounted for by genetics (e_x, e_y) such that

$$X = g_x + e_x$$

$$Y = g_y + e_y$$

The genetic correlation between trait X and trait Y is given by[138]:

$$r_g = \frac{Cov_{g_x,g_y}}{\sqrt{sd(g_x)^2 sd(g_y)^2}}$$

Where r_g Genetic correlation between traits X and Y

Cov_{g_x,g_y} is the covariance of g_x and g_y

$sd(g_x)$ is the standard deviation of g_x

$sd(g_y)$ is the standard deviation of g_y

Traditional methods for computing genetic correlations typically use pedigrees containing individuals with varying degrees of relatedness and diagnosis for both traits of interest. Even in large biobanks, for rare dichotomous traits, it becomes hard to find such datasets. More recent approaches have been implemented to compute genetic correlations if the genotypes are available from cohorts of individuals with both traits of interest[3] and in the absence of that, using summary statistics from genome wide association studies[116]. Estimates of genetic correlation are regularly made in current studies of complex traits, some examples include studies investigating the shared genetic effects between BMI and cognitive function[139], pain, depression and neuroticism[140], Alzheimer's disease and heart function[141].

In this thesis, genetic correlations were computed for voter turnout with a variety of health, socioeconomic, psychiatric and reproductive traits using linkage disequilibrium score regression as implemented using LDHub[4], a centralized repository of summary statistics from genome wide association studies and web interface to compute genetic correlations with a large number of traits.

6.7 SUMMARY STATISTICS BASED MENDELIAN RANDOMIZATION

Mendelian randomization is a technique which uses genetic variants as random instrumental variables in place of an exposure[142] to estimate the causal influence of one risk factor or trait towards a second trait or outcome. The advantages of the choice of genetic variants as instruments are that they remain unchanged over time, they are protective towards the biases introduced by reverse causation, they are plausibly free from ascertainment biases, and after pruning for linkage disequilibrium and in the absence of genetic pleiotropy, they are unlikely to be correlated with other exposure variables[143].

Zhu et. al (2016)[144] extended the mendelian randomization approach to develop a method called summary statistics based mendelian randomization (SMR) which uses genome wide significant SNPs from GWAS summary statistics as instruments and gene expression data from eQTL studies as exposures to identify genes that affect a phenotype at GWAS associated loci in complex traits. They further extended this to develop a new method called generalized summary statistics based mendelian randomization (GSMR)[145] which increases the power of the mendelian randomization approach by leveraging the information from multiple genome wide significant SNPs associated with a trait, pruned for linkage disequilibrium, as instruments to infer causality from correlated traits. They use a method called HEIDI-outlier to remove SNPs showing evidence of pleiotropy towards both the exposure and the outcome. The authors used their approach to demonstrate the causal effect of traits such as blood pressure, serum cholesterol on coronary artery disease, BMI on type 2 diabetes, asthma and osteoarthritis.

According to GSMR, if $z_i = \{z_1, z_2, .. z_m\}$ are independent SNPs after LD pruning, associated with trait x at a threshold surpassing genome wide significance and $\hat{b}_{zx}$ is the effect

of the SNPs z_i on trait x, $\hat{b}_{zy}$ is the effect of the SNPs z_i on trait y, the mediation effect of trait x on trait y, $\hat{b}_{xy} = \frac{\hat{b}_{zy}}{\hat{b}_{zx}}$.

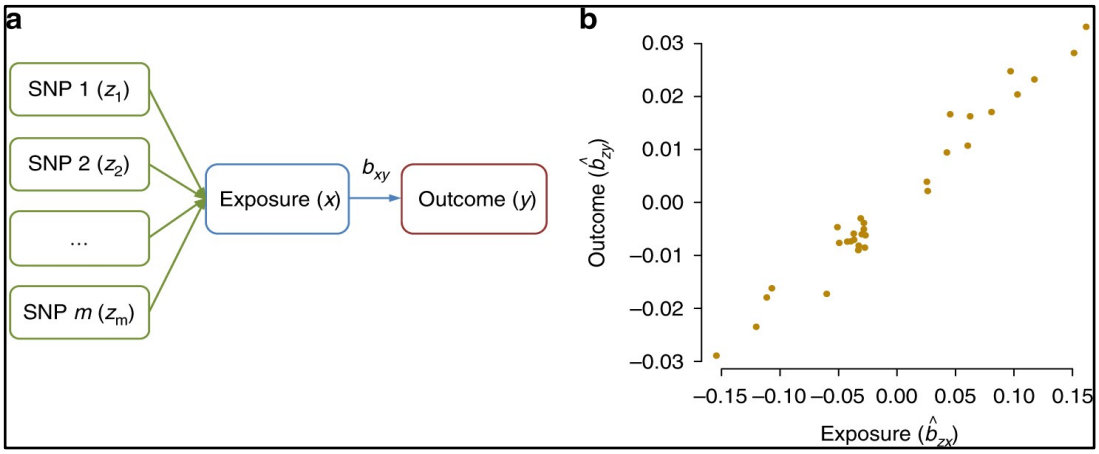


Figure 8. Illustrating GSMR. Panel (**a**) shows SNPs {z_1, z_2 .. z_m } associated with exposure x and b_{xy} is the effect of the exposure x towards outcome y. Panel (**b**) shows that if the exposure x has an effect on outcome y, the effect of z on x, $\hat{b}_{zx}$ is linearly proportional to the effect of z on y, $\hat{b}_{zy}$. ***Figure borrowed from Zhu et. al., Nature Communications 9, Article number: 224 (2018)[139].***

The work presented in this thesis uses the GSMR approach with SNPs found to be associated at a threshold surpassing genome wide significance in genome wide association studies of educational attainment[24] and intelligence test performance[25] as instrumental variables. The increase in voter turnout for each standard deviation increase in genetic predisposition to educational attainment and intelligence test performance as compared to the population mean is demonstrated in both the nationally representative and psychiatric cohorts of iPSYCH2012.

6.8 DESIGN CHOICES FOR THE ANALYSIS OF SECONDARY PHENOTYPES

The first two papers included in this thesis study the genetic basis of suicide attempts in the iPSYCH cohort and the predictive utility of genetic instruments that confer a disposition towards educational attainment and intelligence test performance towards voter turnout. Suicidal behaviour and voter turnout are unascertained traits in the iPSYCH study design and

they are both correlated with the psychiatric disorders[106,146] for which the sample was originally ascertained.

In the first paper studying the genetics of suicide attempts in the iPSYCH cohort, we perform genome wide association studies with and without the main iPSYCH psychiatric diagnoses as covariates and demonstrate the differences introduced by covariate adjustment. We also exploit the comorbidity between affective disorders and suicide attempts to conduct a GWAS comparing individuals with affective disorders and a recorded suicide attempt requiring hospitalization to healthy controls and find a genome wide significant locus in the intronic region of the gene PDE4B.

In the second paper investigating the predictive utility of genetic instruments conferring a predisposition to educational attainment and intelligence test performance towards voter turnout, we take advantage of the case-cohort design of iPSYCH to compare our findings in the nationally representative subset of the iPSYCH2012 cohort to the population of individuals ascertained for severe mental disorders and demonstrate that our findings are consistent between both populations.

7. SUMMARIES OF RESEARCH PAPERS

7.1 Genetics of suicide attempts in individuals with and without mental disorders: A population based genome wide association study

MOTIVATION

Suicidal behaviour has an estimated lifetime prevalence varying between 1.9 and 8.7%[147], accounting for $11.8 billion in lost productivity per year in the United States[148]. Twin studies suggest a heritability of suicidal behaviour varying between 33-51%[149]. The prevalence of suicide attempts is also observed to be higher in individuals with mental disorders[106]. Previous genetic studies of suicidal behaviour have been conducted in small sample sizes of typically a few thousand cases and failed to account for comorbid mental disorders[150]. The case-cohort design of iPSYCH2012 dataset renders the possibility to conduct a genetic study of suicidal behaviour in a larger sample size of individuals with and without mental disorders and further within each of the major mental disorders for which the iPSYCH2012 case sample was ascertained. Since suicidal behaviour is an unascertained secondary trait in iPSYCH2012, which is correlated to the ascertained primary traits, this gives the opportunity to demonstrate the necessity of taking this ascertainment bias into account when conducting such a study.

AIM

- Use complex trait analysis methods to study the genetic basis of suicide attempts in the iPSYCH2012 cohort while demonstrating the need for taking comorbid diagnosis of mental disorders into account.

DATA

The iPSYCH2012[8] dataset, previously described in section 4.1, was utilized in this study. Diagnoses of mental disorders were obtained from the Danish psychiatric central

registers[9]. Cases were individuals in iPSYCH, who had at least one occurrence of non-fatal suicide attempt requiring hospitalization as observed in the Danish psychiatric central register. Controls were individuals with no such occurrence.

METHODS

Quality control of genotypes was performed within a set of individuals of homogeneous genetic ancestry as determined using principal component analysis to exclude SNPs with high missingness, violations of Hardy Weinberg equilibrium in controls, differential missingness between cases and controls, associations to a genotyping wave. Phasing was performed using SHAPEIT3[95] and the genotypes were further imputed to the 1000 genomes phase 3 reference panel using IMPUTE2[124]. Post imputation quality control included exclusion of SNPs with an imputation INFO score < 0.3, minor allele frequency < 0.01, differential imputation quality between cases and controls, associations with an imputation batch. Individuals were excluded for high missingness, relatedness higher than the third degree and deviations from a homogenous genetic origin as determined by principal component analysis.

We used the GCTA[3] software suite to determine estimates of narrow sense SNP heritability of suicidal behaviour in the iPSYCH2012 cohort, with and without adjusting for comorbid mental disorders as covariates. Estimates of SNP heritability were further estimated in individuals with mental disorders and within individuals with autism, schizophrenia, affective disorder and attention-deficit/hyperactivity disorder.

Genome wide associations studies were conducted using the logistic regression model implemented in PLINK[2] comparing individuals with at least one incidence of suicide attempts (N = 6,024) to individuals with no such occurrences (N = 44,240), initially with socio-demographic variables such as age, gender and principal components of genetic ancestry as covariates and further with the addition of diagnosis of mental disorders as covariates. A genome wide association study was also conducted comparing individuals with affective disorders and at least one occurrence of suicide attempt (N = 4,302) to controls with no diagnosis of severe mental disorders (N = 14,935). Characterization of loci reaching genome

wide significance was performed by intersecting them with GeneHancer database of enhancer regions v4.6.0 build 15[146].

IMPORTANT FINDINGS

- SNP heritability of suicide attempt in the iPSYCH2012 cohort was estimated to be 4.6% [95% CI: 2.9 - 6.3%], when adjusting for comorbid mental disorders, this estimate was found to be 1.9% [95% CI: 0.9 - 3.5%].
- SNP heritability of suicide attempt in individuals with severe mental disorders was estimated to be 2.8% [95% CI: 0.7 - 4.9%].
- The SNP heritability of suicide attempt in individuals diagnosed with affective disorders was estimated at 5.6% [95% CI: 1.9 - 9.3%], whereas in individuals diagnosed with autism spectrum disorders, the estimate was 9.6% [1.1 - 18.1].
- Estimates of SNP heritability of suicidal behaviour in individuals without mental disorders, individuals diagnosed with schizophrenia and ADHD was not found to be statistically significant.
- Genome wide association study of suicide attempts in iPSYCH2012 did not reveal any genome wide significant associations when adjusting for socio demographic factors but when diagnosis of mental disorders were introduced as covariates, a significant association was found in an intergenic region on chromosome 20.
- The genome wide association study of individuals with affective disorders and incidence of suicide attempts compared to healthy controls revealed a genome wide significant locus in the intergenic region of PDE4B, a gene observed to be highly expressed in brain related tissues in GTEx[151].

CONCLUSION AND LIMITATIONS

A large genetic study of suicidal behaviour in the iPSYCH2012 cohort revealed the genetic contributions of common single nucleotide polymorphisms to suicidal behaviour. Genome wide association studies revealed novel loci associated with suicide attempts when using comorbid mental disorder diagnoses as covariates, stressing the importance of covariate adjustment as a way to account for ascertainment biases. The limitations of this

study include that by design, the iPSYCH2012 population is young, a population observed to have higher incidence of suicide attempts in Denmark[152] and suicide attempts are underreported in the psychiatric central registers, owing to individuals not seeking hospitalization for such incidents[153].

7.2 Genetic predictors of educational attainment and intelligence test performance predict voter turnout

MOTIVATION

Democracies are built on active electoral participation from their citizens to choose representatives who can advance their views and priorities. Disparities in electoral participation exist between individuals and particularly among those with mental disorders, who might be vulnerable for political exclusion. Twin studies have suggested estimates of the heritability of voter turnout to be between 40-50%[154] whereas social science research has posited a resource based model inclusive of indicators of socio economic status and education as strong predictors of electoral participation[155]. The motivation of this study is to integrate these genetic and social science models of voter turnout by evaluating the predictive utility of genetic instruments that associate with educational attainment and intelligence test performance towards voter turnout in the iPSYCH2012 cohort across three different elections.

AIM

- Using complex trait analysis methods, evaluate the predictive utility of genetic predictors of educational attainment and intelligence test performance towards voter turnout across three elections.
- Compare the portability of this predictive utility between a randomly ascertained, nationally representative subset of the iPSYCH2012 cohort (N = 13,884) to the subset of iPSYCH2012 that was ascertained for individuals with severe mental disorders (N = 33,062).

DATA

The iPSYCH2012 cohort as described in section 4.1 was utilized in this study. The diagnosis of mental disorders was obtained from the Danish psychiatric central registers. Validated voter turnout information was obtained from administrative records at polling stations for three elections in Denmark, including the municipal elections in 2013, European elections in 2014 and the National election in 2015. Genome wide association study summary statistics for educational attainment[27], intelligence test performance[28], height[85] and big five personality traits[29] were obtained from the public domain.

METHODS

The quality control, phasing and imputation of genotypes was performed as previously described in section 7.1 pertaining to paper 1. GCTA[3] was used to estimate narrow sense SNP heritability of voter turnout in each of the three elections in both the nationally representative and psychiatric cohorts of iPSYCH[8]. Polygenic scores for educational attainment and intelligence test performance were calculated using a pruning and thresholding method as implemented in PRSice[136]. Genetic correlations of voter turnout with socio economic, cognitive, education, health and psychiatric traits were calculated using the web interface LDHub[4]. Generalized summary statistics based mendelian randomization was performed using the GSMR[145] module of the GCTA while accounting for linkage disequilibrium within individuals of central European ancestry of the 1000 genomes project phase 3 dataset[59] prior to selection of genome wide significant SNPs as instruments.

IMPORTANT FINDINGS

- Common single nucleotide polymorphisms explained 8 - 10% [SE = 0.0155 - 0.021; $p <= 2.61 \times 10^{-7}$] of the variation in electoral participation in the sub cohort of iPSYCH2012 ascertained for severe mental disorders.
- In the nationally representative sample, the common SNPs explain 10.96% of the variation in voter turnout [SE = 0.044; $p = 6.43 \times 10^{-3}$] in the 2014 European election,

whereas the narrow sense SNP heritability of voter turnout in the higher saliency national and municipal elections was found to be non-significant in this cohort.

- Polygenic scores for educational attainment and intelligence test performance were found to be significantly associated with voter turnout in both the nationally representative and psychiatric sub cohorts with an increase in electoral participation observed for each percentile increment in polygenic scores for these two traits.
- Genetic correlation analysis displayed positive correlations for voter turnout with traits such as educational attainment, childhood IQ, college completion, higher age at child birth, lifespan of the parents and negative correlations with poor health indicators, smoking behaviour and psychiatric disorders.
- Generalized summary statistics based mendelian randomization showed a higher likelihood of electoral participation with each standard deviation increase in predisposition towards educational attainment and intelligence test performance as compared to the population mean.

CONCLUSION AND LIMITATIONS

Complex trait analysis methods integrating the genetic and social science models of voter turnout demonstrated the predictive utility of genetic instruments conferring predisposition to educational attainment and intelligence test performance towards voter turnout. The magnitude of genetic contributions towards voter turnout were shown to be reliant on the saliency of the election. The genetic mechanisms influencing voter turnout were reasonably consistent between a nationally representative population and a population of individuals with severe mental disorders. Despite the sample size, the opportunity to study the genetic influences on voter turnout within each severe mental disorder was limited. While the predictive utility of genetic predictors of educational attainment and intelligence test performance could be demonstrated, these traits are further correlated with personality traits such as conscientiousness which could influence voter turnout but the ability to test such hypotheses was limited owing to the smaller sample sizes used in the genome wide association studies of personality traits.

7.3 Legacy data, whole genome imputation and the analysis of complex traits: Lessons from the iPSYCH Case-Cohort study

MOTIVATION

When faced with the problem of integrating legacy data genotyped using different genotyping arrays with minimal marker overlap, the conventional wisdom has been to phase and impute them separately, perform genome wide association studies using the resulting common markers and then meta-analyze them together to enhance power. However, the accuracy of haplotype estimation and thereby genotype dosages obtained from whole genome imputation increases with increasing sample size of the study and reference cohorts owing to the haplotype estimation tools having more informative haplotypes. Since the samples collected by biobank scale datasets are often much larger than the largest available reference panels available for phasing and imputation and often more informative due to close genetic origins between the study individuals, it makes intuitive sense to find ways to pool together as many samples as possible to get the best possible phasing and imputation accuracies. However, the strategies for data combination in such scenarios has not been well studied. Moreover, the robustness of haplotype estimation tools to the peculiarities in input datasets introduced by data integration protocols involving data from multiple genotyping arrays with low marker overlap has not been well characterized. As phasing and imputation are essential steps in computational pipelines of the data generation process for complex trait analysis, the impact of the choices of data integration protocols on tools used to perform these steps and further their effects on estimates obtained from complex trait analysis needs to be better understood.

AIM

- To investigate the best bioinformatics approaches to integrate legacy data generated over time using different genotyping arrays with minimal marker overlap.
- To benchmark the accuracy of haplotype estimation methods across different scenarios of integrating legacy data.

- To characterize the biases and inaccuracies each choice of haplotype estimation method and data merging strategy could have on the analysis of complex traits.

DATA

The iPSYCH2012, iPSYCH2015, Trios and PGP-UK cohorts as described in section 4.1, were utilized in this study. The diagnosis of mental disorders was obtained from the Danish psychiatric central registers. The haplotype reference consortium HRC(v1.1)[60] dataset is used as the haplotype reference throughout the study. This is the largest reference dataset that can be downloaded onto the secure servers where iPSYCH data is permitted to be stored and analyzed.

METHODS

The consequences arising from four different data integration protocols for two cohorts genotyped using different arrays is investigated. Accuracy of phasing in input datasets with peculiarities introduced by each choice of data integration protocol is compared using three different tools and an approach taking the consensus call at each locus from all three tools across the four aforementioned different data merging protocols is measured using trio offspring whose parental genotypes are known. Accuracy of missing data imputation arising from each choice of data integration protocol and haplotype estimation tool is evaluated at different minor allele frequency bins using 10,000 SNPs masked prior to phasing and 10 whole genome sequenced samples down sampled to each genotyping array. The variation in imputation accuracy is investigated within iPSYCH samples grouped by the birthplace of their parents. The biases arising from each data merging protocol for genome wide association studies is demonstrated by estimating the inflation in summary statistics by performing an association test in unrelated controls of a homogenous genetic ancestry with the genotyping array as the outcome. Simulation of a continuous complex trait with a heritability of 0.5 and each of the 10,000 masked SNPs as causal loci is performed using GCTA[3] to characterize the consequences of choice of data merging protocols in estimates of polygenic scores.

IMPORTANT FINDINGS

- Phasing accuracy relies on the sample size of the target dataset, the density of the study markers and choice of phasing method.
- Missing data imputation is less accurate when performed using haplotype scaffolds generated by data integration protocols that merge data from different genotyping arrays with low marker overlap as compared to phasing and imputing such datasets separately.
- An attenuation in imputation accuracy is observed in samples of non-European origin in the iPSYCH dataset even at common allele frequencies.
- There is an inflation in test statistics when comparing samples genotyped and imputed using different genotyping arrays.
- Polygenic scores calculated using simulated GWAS effect sizes suggest an attenuation in variance explained and a discordance in individuals in the most actionable deciles of polygenic score distribution when PGS is calculated using imputed dosages as compared to true genotypes.
- There is evidence of a bias in imputed dosages towards the major allele in the haplotype reference panel used for imputation and some of the attenuation in PGS performance from imputed dosages as compared to true genotypes can be attributed to this bias.

CONCLUSION AND LIMITATIONS

This empirical study investigating the best approaches towards integrating legacy data shows the variations in accuracy of haplotype estimation and missing data imputation across different data integration protocols. Association analyses with the genotyping array as the outcome show inflated test statistics, suggesting the need to take this bias into account to avoid type I errors in GWAS when comparing samples genotyped and imputed from multiple arrays. Polygenic score analysis using simulated effect sizes shows an attenuation in variance explained by the PGS and individuals dropping out of the most actionable percentiles of risk strata when scores are calculated using imputation dosages as compared to true genotypes.

8. CONCLUSIONS AND FUTURE WORK

The three papers in this thesis demonstrate the opportunities and challenges that come with conducting complex trait analyses in biobank scale datasets. Biobank scale datasets with genetic data from several hundreds of thousands of individuals combined with phenotypic information extracted from electronic health records, participant questionnaires or national health registers have the ability to transform complex trait analysis and bring patient stratification using polygenic scores closer to the clinic. The challenges that come with integrating legacy data, as presented in paper 3 necessitate that bioinformatics be at the forefront of building the infrastructure required for such a shift in clinical practice.

The methods for complex trait analysis are also constantly evolving and since the publication of the first paper in this thesis on the genetics of suicide attempts in the iPSYCH cohort, there have been methods developed to perform cox regression[156] for age of onset of a trait as the outcome, which has shown to offer more power than logistic regressions and new implementations making it scalable for genome wide association studies[157]. While we had to dichotomize suicide attempts as a trait, a cox model treating the time to each suicide attempt requiring hospitalization as the outcome could afford the opportunity to evaluate the phenotypic severity, especially in individuals who have more than one incidence of suicide attempts.

Finally, low pass sequencing of individuals at average coverage around 0.4 - 1x has recently been shown to be a cost effective way of generating genetic data for complex trait analyses without the ancestry biases of traditional genotyping arrays[158–160]. While this may make the challenges in integration of legacy data much easier, the current methods for haplotype estimation and imputation from such datasets are either proprietary or not been benchmarked rigorously. However, the future of complex trait analyses are projected to be using sequencing datasets[161], making fine mapping of causal variants much easier. Once the annotation of noncoding variation becomes more mature, it can pave the way for translating findings from complex trait analyses towards novel biological insights.

9. REFERENCES

1. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
2. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
3. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
4. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
5. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
6. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
7. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
8. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
9. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
10. Pedersen, C. B. The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).
11. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
12. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3**, 779–789 (2002).
13. Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* **211**, 1125–1130 (2019).
14. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).

15. Weight, S. A Frequency Curve Adapted to Variation in Percentage Occurrence. *J. Am. Stat. Assoc.* **21**, 162–178 (1926).
16. Dempster, E. R. & Lerner, I. M. Heritability of Threshold Characters. *Genetics* **35**, 212–236 (1950).
17. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
18. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
19. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
20. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
21. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
22. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
23. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
24. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
25. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
26. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
27. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
28. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
29. Lo, M.-T. *et al.* Genome-wide analyses for personality traits identify six genomic loci and

show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2017).

30. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).
31. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
32. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
33. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
34. Strittmatter, W. J. & Roses, A. D. Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* **19**, 53–77 (1996).
35. Duggirala, R. *et al.* Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *Am. J. Hum. Genet.* **64**, 1127–1140 (1999).
36. Ghosh, S. *et al.* Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2198–2203 (1999).
37. Aouizerat, B. E. *et al.* A genome scan for familial combined hyperlipidemia reveals evidence of linkage with a locus on chromosome 11. *Am. J. Hum. Genet.* **65**, 397–412 (1999).
38. Samani, N. J. *et al.* A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. *Am. J. Hum. Genet.* **77**, 1011–1020 (2005).
39. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493–507 (1999).
40. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
41. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
42. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
43. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

44. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
45. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
46. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
47. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
48. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
49. Hrdlickova, B., de Almeida, R. C., Borek, Z. & Withoff, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta* **1842**, 1910–1922 (2014).
50. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
51. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
52. Hill, W. G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, 615–628 (1968).
53. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
54. Ding, K. & Kullo, I. J. Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. *Eur. J. Hum. Genet.* **15**, 228–236 (2007).
55. Stram, D. O. Tag SNP selection for association studies. *Genet. Epidemiol.* **27**, 365–374 (2004).
56. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
57. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).

58. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
59. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
60. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
61. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
62. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
63. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
64. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
65. Macé, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat. Commun.* **8**, 744 (2017).
66. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
67. Dajani, R. *et al.* CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations. *Sci. Rep.* **5**, 13391 (2015).
68. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–92 (2014).
69. Satterstrom, F. K. *et al.* Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
70. Singh, T. *et al.* Rare schizophrenia risk variants are enriched in genes shared with neurodevelopmental disorders. *Cold Spring Harbor Laboratory* 069344 (2016) doi:10.1101/069344.
71. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study

designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

72. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).
73. Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120**, 571–580 (2006).
74. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
75. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
76. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
77. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
78. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).
79. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
80. Hansen, T. F. *et al.* DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open* **9**, e028401 (2019).
81. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
82. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
83. Davis, L. Psychiatric Genomics, Phenomics, and Ethics Research In A 270,000-Person Biobank (BioVU). *Eur. Neuropsychopharmacol.* **29**, S739–S740 (2019).
84. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
85. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

86. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
87. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
88. Schork, A. J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
89. Shen, X. *et al.* A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. *Nat. Commun.* **11**, 2301 (2020).
90. Zeggini, E. & Ioannidis, J. P. A. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191–201 (2009).
91. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
92. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
93. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
94. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
95. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
96. Sinnott, J. A. & Kraft, P. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum. Genet.* **131**, 111–119 (2012).
97. Johnson, E. O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum. Genet.* **132**, 509–522 (2013).
98. Li, H., Gail, M. H., Berndt, S. & Chatterjee, N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet. Epidemiol.* **34**, 427–433 (2010).
99. Schifano, E. D., Li, L., Christiani, D. C. & Lin, X. Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Genet.* **92**, 744–759 (2013).
100. Swanson, J. M. The UK Biobank and selection bias. *The Lancet* vol. 380 110 (2012).

101. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
102. Day, F. R., Ong, K. K. & Perry, J. R. B. Elucidating the genetic basis of social interaction and isolation. *Nat. Commun.* **9**, 2457 (2018).
103. Lane, J. M. *et al.* Biological and clinical insights from genetics of insomnia symptoms. *Nat. Genet.* **51**, 387–393 (2019).
104. Meier, S. M. *et al.* Genetic Variants Associated With Anxiety and Stress-Related Disorders: A Genome-Wide Association Study and Mouse-Model Study. *JAMA Psychiatry* **76**, 924–932 (2019).
105. Demontis, D. *et al.* Genome-wide association study implicates CHRNA2 in cannabis use disorder. *Nat. Neurosci.* **22**, 1066–1074 (2019).
106. Brådvik, L. Suicide Risk and Mental Disorders. *Int. J. Environ. Res. Public Health* **15**, (2018).
107. Bandelow, B. & Michaelis, S. Epidemiology of anxiety disorders in the 21st century. *Dialogues Clin. Neurosci.* **17**, 327–335 (2015).
108. Green, B., Young, R. & Kavanagh, D. Cannabis use and misuse prevalence among people with psychosis. *Br. J. Psychiatry* **187**, 306–313 (2005).
109. Skalisky, J. *et al.* Prevalence and Correlates of Cannabis Use in Outpatients with Serious Mental Illness Receiving Treatment for Alcohol Use Disorders. *Cannabis Cannabinoid Res* **2**, 133–138 (2017).
110. Richardson, D. B., Rzehak, P., Klenk, J. & Weiland, S. K. Analyses of case-control data for additional outcomes. *Epidemiology* **18**, 441–445 (2007).
111. Monsees, G. M., Tamimi, R. M. & Kraft, P. Genome-wide association scans for secondary traits using case-control samples. *Genet. Epidemiol.* **33**, 717–728 (2009).
112. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* **8**, e1003032 (2012).
113. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
114. Chervova, O. *et al.* The Personal Genome Project-UK, an open access resource of human multi-omics data. *Sci Data* **6**, 257 (2019).
115. Schmidt, M., Pedersen, L. & Sørensen, H. T. The Danish Civil Registration System as a tool in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).

116. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
117. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
118. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
119. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
120. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* vol. 3 160025 (2016).
121. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
122. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *Cold Spring Harbor Laboratory* 797944 (2020) doi:10.1101/797944.
123. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
124. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
125. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
126. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
127. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
128. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
129. Choi, S. W. & O'Reilly, P. SA20 - PRSice 2: POLYGENIC RISK SCORE SOFTWARE (UPDATED) AND ITS APPLICATION TO CROSS-TRAIT ANALYSES. *Eur. Neuropsychopharmacol.* **29**, S832 (2019).

130. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
131. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
132. Marquez-Luna, C., Gazal, S., Loh, P. R., Kim, S. S. & Furlotte, N. LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* (2020).
133. Ni, G. *et al.* A comprehensive evaluation of polygenic score methods across cohorts in psychiatric disorders. *Genetic and Genomic Medicine* (2020) doi:10.1101/2020.09.10.20192310.
134. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).
135. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
136. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
137. Becker, K. G. The common variants/multiple disease hypothesis of common complex genetic disorders. *Med. Hypotheses* **62**, 309–317 (2004).
138. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
139. Marioni, R. E. *et al.* Assessing the genetic overlap between BMI and cognitive function. *Mol. Psychiatry* **21**, 1477–1482 (2016).
140. Meng, W. *et al.* Genetic correlations between pain phenotypes and depression and neuroticism. *Eur. J. Hum. Genet.* **28**, 358–366 (2020).
141. Sáez, M. E. *et al.* Genome Wide Meta-Analysis identifies common genetic signatures shared by heart function and Alzheimer's disease. *Sci. Rep.* **9**, 16665 (2019).
142. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
143. Smith, G. D. & Ebrahim, S. *Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies*. (National Academies Press (US), 2008).

144. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
145. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
146. Kelly, B. D. & Nash, M. Voter participation among people attending mental health services in Ireland. *Ir. J. Med. Sci.* **188**, 925–929 (2019).
147. Nock, M. K. *et al.* Suicide and suicidal behavior. *Epidemiol. Rev.* **30**, 133–154 (2008).
148. Reducing suicide: A national imperative. **496**, (2002).
149. Statham, D. J. *et al.* Suicidal behaviour: an epidemiological and genetic study. *Psychol. Med.* **28**, 839–855 (1998).
150. Galfalvy, H. *et al.* A genome-wide association study of suicidal behavior. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168**, 557–563 (2015).
151. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
152. Reuter Morthorst, B., Soegaard, B., Nordentoft, M. & Erlangsen, A. Incidence Rates of Deliberate Self-Harm in Denmark 1994-2011. *Crisis* **37**, 256–264 (2016).
153. Hawton, K., Saunders, K. E. A. & O'Connor, R. C. Self-harm and suicide in adolescents. *Lancet* **379**, 2373–2382 (2012).
154. Loewen, P. J. & Dawes, C. T. The Heritability of Duty and Voter Turnout. *Polit. Psychol.* **33**, 363–373 (2012).
155. Brady, H. E., Verba, S. & Schlozman, K. L. Beyond Ses: A Resource Model of Political Participation. *Am. Polit. Sci. Rev.* **89**, 271–294 (1995).
156. Hughey, J. J. *et al.* Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics* **20**, 805 (2019).
157. He, L. & Kulminski, A. M. Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. *Genetics* **215**, 41–58 (2020).
158. Martin, A. R. *et al.* Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Cold Spring Harbor Laboratory* 2020.04.27.064832 (2020) doi:10.1101/2020.04.27.064832.
159. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *Cold Spring Harbor Laboratory* 632141 (2019) doi:10.1101/632141.

160. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Cold Spring Harbor Laboratory* 2020.04.29.068452 (2020) doi:10.1101/2020.04.29.068452.
161. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

10. APPENDICES

PAPER 1

Article

Supplementary Information

Declaration of Co-authorship

PAPER 2

Article

Supplementary Information

Supplementary Tables

Declaration of Co-Authorship

PAPER 3

Manuscript

Supplementary Information

Supplementary Tables

Declaration of Co-Authorship

PAPER 1

Erlangsen A[a], **Appadurai V[a]**, Wang Y, Turecki G, Mors O, Werge T, Mortensen PB, Starnawska A, Børglum AD, Schork A, Nudel R, Bækvad-Hansen M, Bybjerg-Grauholm J, Hougaard DM, Thompson WK, Nordentoft M[b], Agerbo E[b]. ***Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study***. Mol Psychiatry. 2020 Oct;25(10):2410-2421. doi: 10.1038/s41380-018-0218-y. Epub 2018 Aug 16. PMID: 30116032; PMCID: PMC7515833.

[a] Shared first authorship
[b] Shared last authorship

ARTICLE

SUPPLEMENTARY INFORMATION

DECLARATION OF CO-AUTHORSHIP

Molecular Psychiatry
https://doi.org/10.1038/s41380-018-0218-y

ARTICLE

Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study

Annette Erlangsen[1,2,3,4] · Vivek Appadurai[1,5] · Yunpeng Wang[1,6,7] · Gustavo Turecki[8] · Ole Mors[1,9] · Thomas Werge[1,5,10] · Preben B. Mortensen[1,11] · Anna Starnawska[1,12] · Anders D. Børglum[1,12,13] · Andrew Schork[1,5] · Ron Nudel[1,5] · Marie Bækvad-Hansen[1,14] · Jonas Bybjerg-Grauholm[1,14] · David M. Hougaard[1,14] · Wesley K. Thompson[1,6,7,9,15] · Merete Nordentoft[1,2,10,16] · Esben Agerbo[1,11]

Received: 13 February 2018 / Revised: 14 May 2018 / Accepted: 4 June 2018

Abstract

Family studies have shown an aggregation of suicidal behavior in families. Yet, molecular studies are needed to identify loci accounting for genetic heritability. We conducted a genome-wide association study and estimated single nucleotide polymorphisms (SNP) heritability for a suicide attempt. In a case-cohort study, national data on all individuals born in Denmark after 1981 and diagnosed with severe mental disorders prior to 2013 ($n = 57{,}377$) and individuals from the general population ($n = 30{,}000$) were obtained. After quality control, the sample consisted of 6024 cases with an incidence of suicide attempt and 44,240 controls with no record of a suicide attempt. Suggestive associations between SNPs, rs6880062 (p-value: 5.4×10^{-8}) and rs6880461 (p-value: 9.5×10^{-8}), and suicide attempt were identified when adjusting for socio-demographics. Adjusting for mental disorders, three significant associations, all on chromosome 20, were identified: rs4809706 (p-value: 2.8×10^{-8}), rs4810824 (p-value: 3.5×10^{-8}), and rs6019297 (p-value: 4.7×10^{8}). Sub-group analysis of cases with affective disorders revealed SNPs associated with suicide attempts when compared to the general population for gene PDE4B. All SNPs explained 4.6% [CI-95: 2.9–6.3%] of the variation in suicide attempt. Controlling for mental disorders reduced the heritability to 1.9% [CI-95: 0.3–3.5%]. Affective and autism spectrum disorders exhibited a SNP heritability of 5.6% [CI-95: 1.9–9.3%] and 9.6% [CI-95: 1.1–18.1%], respectively. Using the largest sample to date, we identified significant SNP associations with suicide attempts and support for a genetic transmission of suicide attempt, which might not solely be explained by mental disorders.

Introduction

The lifetime prevalence of suicide attempts is estimated to be 2.7% [1, 2]. Family studies using clinical and epidemiological data have consistently shown an aggregation of suicidal behavior in families [3–7]. Both twin and adoption studies have indicated that genes account for as much as 30–50% of the observed familial aggregation [8–10]. Hence, molecular studies have attempted to identify specific genes that contribute to suicide risk [11].

The advantage of molecular genetic studies lies in the estimation of genetics effects via single nucleotide polymorphisms (SNPs) rather than through familial relationships. Candidate gene and genome-wide association study (GWAS) designs have identified numerous loci associated with suicide attempts, including *5HTR2A* (rs1885884) [12]; a locus on 2p25 (rs300774) [11, 13]; and *ABI3BP* (rs2576377) [14] among others [15]. Still, to date most molecular genetic studies, even combined ones, have been conducted on small sample sizes (sample < 8900; cases < 2810) [11, 16], implying that they potentially were underpowered and subject to false positives as well as false negatives. Additionally, studies have rarely adjusted for mental disorders or included a non-psychiatric, population-representative sample.

Shared first authorship: Annette Erlangsen, Vivek Appadurai.

Shared last authorship: Merete Nordentoft, Esben Agerbo.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41380-018-0218-y) contains supplementary material, which is available to authorized users.

✉ Annette Erlangsen
Annette.Erlangsen@regionh.dk

Extended author information available on the last page of the article

Published online: 16 August 2018

Mental disorders are strongly linked to suicidal behavior [17–19]. Indeed, often mental disorders and suicidality are seen as clinically related, e.g., suicidal ideation is considered as a qualifying symptom for the diagnosis of depression. Given that liability for a wide range of mental disorders is associated with a common set of genetic factors [20, 21], it is plausible that the effect of genes on suicidal behavior could be mediated through their impact on mental disorders [11, 22, 23]. Although family studies have indicated support for a generic transmission of suicidal behavior independent of mental disorders [3, 4, 6, 10, 22, 23], this remains to be confirmed using molecular genetic data.

The aim of this study was to examine whether specific genetic variants are associated with suicide attempt in a population-based sample of individuals with and without severe mental disorders using a GWAS approach while controlling for mental disorders. In addition, we aimed to estimate the heritability explained by common genetic variation while taking mental disorders into account.

Material and methods

Design and data source

We used a nested case-cohort study design [24]. In Denmark, a unique personal id-number is assigned to all persons born or migrating into the country [25]. This id-number was introduced in 1968 and is listed in the Danish Civil Registration System along with data on gender, birth date, and parents' id number [26]. Since 1981, blood samples from all newborns in Denmark have been archived in the Danish Newborn Screening Biobank at the Danish Statens Serum Institut [27, 28]. Likewise, individual records of all hospital contacts for mental and somatic disorders have been recorded in the Danish Psychiatric Central Research Register since 1968 and the National Registry of Patients since 1977, respectively [29, 30].

Sample

The iPSYCH dataset was constructed from complete and consecutive birth cohorts of all singletons born in Denmark between May 1, 1981 and December 31, 2005 [31]. Among individuals who were residents on their first birthday ($N = 1{,}472{,}762$ persons), those who later were diagnosed with one or more severe mental disorders prior to 31st December 2012 ($n = 57{,}377$) were selected. Severe mental disorders were identified as the following diagnoses recorded according to the 10th revision of International Classification of Diseases (ICD-10) [32]: schizophrenia (F20), bipolar disorders (F30, F31), affective disorders (F30–39), autism spectrum disorders (F84.0, F84.1, F84.5, F84.8, or F84.9), anorexia (F50), and ADHD (F90.0). In addition, a population-based random sample of individuals ($n = 30{,}000$) was included in iPSYCH [31].

Cases

Individuals in the iPSYCH sample who at some point prior to December 31, 2012 had been recorded with one or more incidents of non-fatal suicide attempts were considered as cases. Suicide attempts were identified by screening the Danish Psychiatric Central Research Register and the National Registry of Patients for diagnoses of suicide attempts (ICD-10: X60–X84). In addition, contacts where the "reasons for contact"-variable indicated suicide attempt were included, as well as combinations of diagnoses where the main diagnosis had been recorded as a mental disorder (ICD-10: F chapter) together with a diagnosis of poisoning by drugs or other substances (ICD-10: T36–T50, T52–T60) or injuries to hand, wrist, and forearm (ICD-10: S51, S55, S59, S61, S65, S69). This is a well-established proxy for suicide attempt [33, 34].

Controls

All persons who had not been recorded with one or more suicide attempts were included in the control group. The control group consisted of persons with mental disorders as well as persons with no mental disorders.

Extraction and genotyping

The Danish Newborn Screening Biobank consists of blood spot samples from almost 100% of all newborns in Denmark, which are collected 4–7 days after birth through heel prick and stored at −20 °C [27, 28, 35]. Two 3.2 mm disks were punched from blood spots for each individual. Genomic DNA was extracted and whole genome amplified (WGA) with the use of Extract-N-Amp Blood PCR Kit (Sigma-Aldrich, Seelze, Germany) and RepliG kit (Qiagen, Venlo, The Netherlands) [35, 36]. Genotyping was carried out using Infinium PsychChip v1.0 array (Illumina, San Diego, CA, USA) according to the manufacturer's instructions and handled by the Broad Institute (Boston, MA, USA) over 23 waves. The extraction and genotyping procedures have previously been tested by comparing call and conflict rates of WGA DNA from >20-year-old stored blood spots with those of high quality, recent genomic DNA from the same individuals; concluding a high quality of the WGA DNA [35, 37].

Genotype quality control and imputation

After initial genotyping and quality control at the Broad Institute, a total of 77,639 subjects and 554,360 SNPs were

obtained. A subset of 246,369 high-quality SNPs was identified by excluding SNPs showing deviations from the Hardy–Weinberg equilibrium (p-value $< 1 \times 10^{-6}$) in controls, minor allele frequencies <1%, multi-allelic SNPs, and SNPs in non-autosomal loci within a subset of individuals with a homogenous European ancestry. This set of high-quality SNPs were pre-phased using SHAPEIT3 [38] and the resulting haplotypes were imputed in 10 batches using IMPUTE2 [39] and the 1000 genomes phase3 reference haplotypes, which yielded a total of 80.7 million variants.

The post-imputation quality control excluded SNPs for the following, non-exclusive reasons: variants with minor allele frequencies < 0.001 ($n \approx 67$ million); variants with imputation INFO scores < 0.2 ($n \approx 17$ million); SNPs missing in more than 10% of the imputed samples ($n \approx 1.8$ million); SNPs showing strong associations to genotyping wave ($n \approx 291{,}937$) or imputation batch ($p < 5 \times 10^{-8}$) ($n = 33$); SNPs showing differential imputation quality between psychiatric cases and controls ($p < 1 \times 10^{-6}$) ($n \approx 527{,}912$); and SNPs violating Hardy–Weinberg equilibrium in controls. Finally, a total of 11,601,089 markers were retained.

Population stratification and kinship

Principal components of genetic ancestry were generated using Eigensoft smartPCA [40]. Using a subset of 47,856 individuals whose parents and grandparents were born in Denmark as our reference, we censored individuals deviating from the multivariate mean of the joint distribution of first 10 principal components. Principal components were re-generated using the remaining samples and the second round of censoring was performed. KING was used to generate kinship coefficients to ensure that no two samples were related beyond the second degree [41]. The imputation and quality control process has previously been documented [42].

Statistical analysis

The GWAS was conducted by comparing the imputed additive genotype dosages between cases with one or more suicide attempts to controls with no recorded suicide attempt. Logistic regressions were applied to calculate log odds ratios and 95% confidence intervals using PLINK2 [43] and the R package qqman [44] to generate Manhattan plots. The level of significance and suggestive significance was set to $p < 5 \times 10^{-8}$ and $p < 5 \times 10^{-6}$, respectively. Two different GWAS models were employed; Model 1 contained covariates on gender, years under follow-up where the participant was 15 years of age or older, and the first 10 principal components of genetic ancestry. Model 2: In addition to the covariates used in Model 1, binary covariates for diagnosis of schizophrenia, bipolar disorders, affective disorders, autism spectrum disorders, anorexia, and any other disorder (ICD-10: F chapter) were included. In a third association study, we compared 4302 individuals with affective disorder and at least one recorded suicide attempt to 14,938 controls with affective disorders and no suicide attempt. However, a lack of association was attributed to low statistical power and the model was modified by using only population-based sample of healthy controls ($n = 13{,}294$) while adjusting for socio-demographic conditions.

The GREML approach implemented in the genome-wide complex trait analysis (GCTA) software package was utilized to estimate the SNP variation [45]. A genetic relatedness cutoff of 0.034 was applied to account for cryptic relatedness between samples. The analysis was adjusted for the same covariates as Model 1.

Enhancer annotations

Enhancer annotations were taken from the GeneHancer database [46], and enhancer IDs corresponded to those in GeneCards (v4.6.0 Build 15). SNP positions were converted to genome build hg38 using the UCSC liftOver [47] and were subsequently mapped onto enhancers using BEDTools [48].

Ethical permission

Anonymized data were stored in the Computerome database at the Technical University of Denmark with restricted access. The project has been approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, and Danish Newborn Screening Biobank Steering Committee.

Results

In all, 77,639 individuals born in Denmark between May 1, 1981 and December 31, 2005 were genotyped. Of these, 12,128 participants were excluded during the quality control and 15,247 did not reach the age of 15 during follow-up (Supplementary Figure 1).

The sample consisted of 50,264 individuals over the age of 15 years on December 31, 2012. Of these, 6024 (12.0%) persons had been recorded with a suicide attempt while the remaining 44,240 (88.0%) were considered as controls (Supplementary Table 1).

The results of the GWAS for Model 1 (genomic inflation factor, $\lambda_{gc} = 1.04$) showed suggestive associations on chromosomes 1, 19, and 20. Furthermore, associations for SNPs rs6880062 (NC_000005.9:g.153298025A>G; p-value: 5.4×10^{-8}) and rs6880461 (NC_000005.9:g.153298024G>A; p-value: 9.5×10^{-8}) on chromosome 5

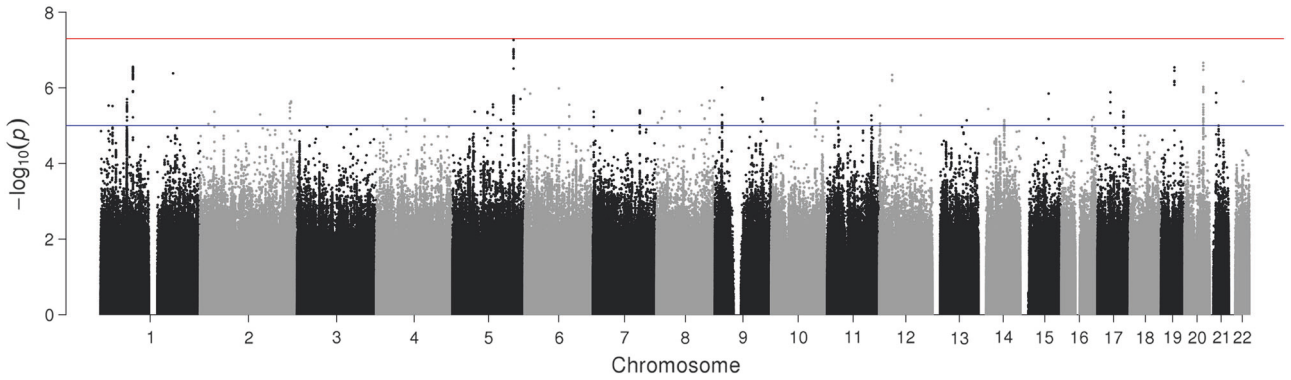


Fig. 1 Manhattan plot of genome-wide association study (GWAS) of suicide attempt adjusted for: gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry (Model 1). *Note*: The red line marks the threshold of $p = 5 \times 10^{-8}$ for significant associations while the blue line marks the threshold of $p = 5 \times 10^{-6}$ for suggestive associations

fell short of genome-wide significance, as seen in Fig. 1 and Table 1.

GWAS Model 2 (genomic inflation factor, $\lambda_{gc} = 1.02$), revealed three genome-wide significant SNPs; rs4809706 (NC_000020.10:g.47193719A>G; *p*-value: 2.8×10^{-8}), rs4810824 (NC_000020.10:47193708G>C; *p*-value: 3.5×10^{-8}), and rs6019297 (NC_000020.10:g.47197230C>T; *p*-value: 4.7×10^{-8}) on chromosome 20 (Fig. 2). In addition, suggestive associations were noted on chromosomes 6, 8, and 9. The three significant SNPs showed suggestive associations in Model 1.

The third association study comparing individuals with depressive disorder and a suicide attempt to healthy controls (genomic inflation factor $\lambda_{gc} = 1.08$) revealed a genome-wide significant locus on chromosome 1 with the lead SNP rs4554696 (NC_000001.10:g.66408011C>T; *p*-value: 5.63×10^{-9} OR $= 0.85 \pm 0.05$ 95% CI) and suggestive associations on chromosomes 2, 5, and 9 (Fig. 3). The Manhattan plot for the GWAS and a GTEX expression plot [49] for PDE4B are presented in Supplementary Figure 2–3 and Supplementary Table 2.

The SNP heritability of suicide attempt was found to be 4.6% [CI-95: 2.9–6.3%] when adjusting for socio-demographic characteristics (Table 2). When introducing diagnosed mental disorders as binary covariates in the GCTA model, SNP heritability was reduced to 1.9% [CI-95: 0.3–3.5%]. Calculation of SNP heritability among people with no mental disorders did not render a significant estimate [OR = 3.3%; CI-95: −0.1 to 7.6%], while a significant h^2 estimate of 2.8% [0.7–4.9%] was noted within subjects with mental disorders. The SNP heritability was further estimated within each subgroup of mental disorders while adjusting for socio-demographic covariates. Individuals with affective disorders were estimated to carry a significant SNP heritability of 5.6% [1.9–9.3%] while those with autism spectrum disorders had 9.6% [1.1–18.1%]. The SNP heritability for individuals with schizophrenia and ADHD were 0% [−27.5 to 27.5%] and 0.4% [−6.6 to 7.4%], respectively.

Discussion

To our knowledge, this is the largest molecular genetic study to examine suicide attempt. By including a population-based sample, it was possible to adjust for the presence of mental disorders, hereby, addressing the possible mediating role of mental disorders. Albeit significant associations were identified by this study, the heritability estimates suggest that diagnosed mental disorders do not fully explain the genetic transmission of a suicide attempt.

GWAS

The lead SNP, rs6880062 on chromosome 5, of Model 1 is intergenic, with the closest gene being *FAM114A2* (distance = 71.6 kb) and has not previously been associated with suicidal behavior [15]. When intersected with GeneHancer database of enhancer regions (version: 4.6.0 build 15) [46], one of the suggestive SNPs on chromosome 22, rs150801052 (*p*-value: 6.78×10^{-7}) was found to be within an enhancer (Enhancer ID: GH22H035858) for the protein-coding gene myoglobin, elevated levels of which has previously been suggested for monitoring in two previous cases involving an adolescent female with suicide attempt and an older adult with depression and psychosis [50, 51].

The enhancer was also connected to the *RBFOX2* gene, which encodes an RNA-binding protein that is thought to be involved in exon splicing in the nervous system. Recently,

Table 1 SNP showing significant or suggestive association with a suicide attempt in adjusted GWAS

Model	SNP	Chromosome	Position	A1/A2	Minor allele frequency	INFO	Odds ratio [CI-95%][a]	*p*-Value
Model 1: Adjusted for socio-demographic conditions[b]								
	rs12130410	1	81359237	C/A	0.0593	0.963265	0.7948 ± 0.089964	5.67E−07
	rs12118384	1	81360791	G/A	0.0607	0.963841	0.7935 ± 0.088984	3.55E−07
	rs72134670	1	81361781	AGTGT/A	0.0592	0.970028	0.7956 ± 0.089768	5.90E−07
	rs114428613	1	81363409	T/A	0.0592	0.971558	0.7945 ± 0.089768	5.05E−07
	rs12122457	1	81364694	G/C	0.0592	0.972986	0.7941 ± 0.089768	4.77E−07
	rs12119932	1	81364728	T/G	0.0592	0.973008	0.7941 ± 0.089768	4.76E−07
	rs12119912	1	81364903	A/G	0.0592	0.973055	0.7941 ± 0.089768	4.74E−07
	rs201713449	1	81365218	G/GCA	0.0592	0.973134	0.7940 ± 0.089768	4.72E−07
	rs116584414	1	81365419	T/G	0.0592	0.973185	0.7940 ± 0.089768	4.69E−07
	rs79479721	1	81365444	T/G	0.0592	0.973191	0.7940 ± 0.089768	4.69E−07
	rs115355703	1	81365722	A/C	0.0592	0.973267	0.7939 ± 0.089768	4.66E−07
	rs114035422	1	81366040	A/G	0.0592	0.973354	0.7939 ± 0.089768	4.64E−07
	rs115194926	1	81366100	G/T	0.0592	0.973373	0.7939 ± 0.089768	4.63E−07
	rs115255447	1	81366951	T/A	0.0592	0.97427	0.7935 ± 0.089768	4.38E−07
	rs115410239	1	81369030	T/A	0.0592	0.977419	0.7937 ± 0.089572	4.30E−07
	rs12132818	1	81371930	C/G	0.0592	0.98065	0.7919 ± 0.089572	3.23E−07
	rs72940613	1	81373376	A/G	0.0595	0.976727	0.7927 ± 0.089376	3.55E−07
	rs6681193	1	81374082	C/T	0.0592	0.982108	0.7913 ± 0.089572	2.94E−07
	rs114462636	1	81375885	C/T	0.0589	0.985097	0.7916 ± 0.089572	3.20E−07
	rs12130052	1	81379254	T/A	0.0588	0.989481	0.7932 ± 0.089376	3.80E−07
	rs12130531	1	81380084	C/T	0.0588	0.989333	0.7932 ± 0.089376	3.77E−07
	rs12121366	1	81381621	T/C	0.0589	0.990004	0.7923 ± 0.089376	3.26E−07
	rs112409531	1	81384412	A/G	0.0588	0.990039	0.7933 ± 0.089376	3.80E−07
	rs12135658	1	81384847	C/G	0.0594	0.991295	0.7967 ± 0.088592	5.08E−07
	rs17105681	1	81385360	C/G	0.0596	0.991544	0.7974 ± 0.088592	5.37E−07
	rs12125794	1	81386101	A/C	0.0595	0.991884	0.7974 ± 0.088592	5.49E−07
	rs72940624	1	81387664	T/G	0.0595	0.992104	0.7972 ± 0.088592	5.36E−07
	rs74092049	1	81387673	T/C	0.0595	0.992106	0.7972 ± 0.088592	5.35E−07
	rs74092050	1	81387892	A/G	0.0596	0.99158	0.7971 ± 0.088592	5.12E−07
	rs17105699	1	81391541	A/G	0.0589	0.991587	0.7928 ± 0.08918	3.36E−07
	rs74092063	1	81393160	G/A	0.059	0.991599	0.7923 ± 0.08918	3.05E−07
	rs7548691	1	81396755	G/A	0.0591	0.992042	0.7920 ± 0.08918	2.91E−07
	rs17105728	1	81398266	T/C	0.0589	0.992018	0.7917 ± 0.08918	2.89E−07
	rs12133801	1	81401114	G/A	0.0589	0.993456	0.7919 ± 0.08918	2.89E−07
	rs79780846	1	81419288	G/A	0.0588	0.987839	0.7911 ± 0.089572	2.83E−07
	rs12139190	1	81420151	A/C	0.0587	0.987233	0.7913 ± 0.089572	3.05E−07
	rs72940689	1	81428767	G/A	0.0586	0.978883	0.7897 ± 0.09016	2.78E−07
	rs2560038	5	153272968	G/A	0.3232	0.993039	0.8960 ± 0.04214	3.10E−07
	rs1593827	5	153280893	G/A	0.3112	0.995496	0.8914 ± 0.042532	1.17E−07
	rs898709	5	153286964	A/G	0.3117	1	0.8929 ± 0.042336	1.66E−07
	rs1870738	5	153287549	T/A	0.3126	0.998605	0.8921 ± 0.042336	1.31E−07
	rs2085865	5	153290253	G/A	0.3121	0.998881	0.8913 ± 0.042336	1.06E−07
	rs2035099	5	153290835	C/A	0.3121	0.998768	0.8914 ± 0.042336	1.07E−07
	rs7726518	5	153293565	G/A	0.3123	0.998245	0.8914 ± 0.042336	1.06E−07
	rs2126160	5	153294271	C/T	0.3123	0.99833	0.8914 ± 0.042336	1.07E−07
	rs7732031	5	153294900	A/G	0.3123	0.998254	0.8914 ± 0.042336	1.08E−07

Table 1 (continued)

Model SNP	Chromosome	Position	A1/A2	Minor allele frequency	INFO	Odds ratio [CI-95%][a]	*p*-Value
rs1599408	5	153296017	C/T	0.3125	0.998001	0.8927 ± 0.042336	1.53E−07
rs1599409	5	153296064	C/A	0.3123	0.998137	0.8915 ± 0.042336	1.10E−07
rs1599410	5	153296073	C/T	0.3123	0.998137	0.8915 ± 0.042336	1.10E−07
rs6880461	5	153298024	G/A	0.3138	0.997715	0.8911 ± 0.042336	9.52E−08
rs6880062	5	153298025	A/G	0.3107	0.989144	0.8884 ± 0.042728	5.44E−08
rs2614123	5	153300215	C/T	0.3139	0.996726	0.8914 ± 0.042336	1.06E−07
rs7862648	9	18290857	A/G	0.2187	0.991736	1.1212 ± 0.045864	9.80E−07
rs112595860	12	32640591	C/G	0.2172	0.931883	0.8798 ± 0.049784	4.55E−07
rs113067218	12	32641482	CT/C	0.2174	0.932194	0.8814 ± 0.049784	6.54E−07
rs10506086	12	32642659	C/T	0.2162	0.930623	0.8808 ± 0.04998	6.17E−07
rs4810824	20	47193708	G/C	0.6268	0.959962	0.8994 ± 0.040376	2.65E−07
rs4809706	20	47193719	A/G	0.6273	0.958887	0.8987 ± 0.040376	2.19E−07
rs6125386	20	47196367	A/G	0.6323	0.971916	0.9042 ± 0.04018	9.42E−07
rs6019297	20	47197230	C/T	0.625	0.958847	0.9002 ± 0.040376	3.34E−07
rs150801052	22	36255928	AT/A	0.0167	0.677581	1.5076 ± 0.161896	6.78E−07
Model 2: Adjusted for socio-demographic conditions and mental disorders[c]							
rs4053798	6	85943189	A/C	0.9945	0.926872	0.5418 ± 0.242844	7.60E−07
rs76426299	8	58350711	G/A	0.0189	0.556548	1.6189 ± 0.191688	8.40E−07
rs7862648	9	18290857	A/G	0.2187	0.991736	1.1281 ± 0.047824	7.82E−07
rs4810824	20	47193708	G/C	0.6268	0.959962	0.8884 ± 0.04214	3.56E−08
rs4809706	20	47193719	A/G	0.6273	0.958887	0.8875 ± 0.04214	2.80E−08
rs6019294	20	47194199	G/A	0.5818	0.974655	0.9014 ± 0.04116	7.68E−07
rs6066764	20	47194229	C/A	0.5759	0.971501	0.8994 ± 0.04116	4.57E−07
rs6066765	20	47194232	G/C	0.5759	0.971509	0.8994 ± 0.04116	4.57E−07
rs6019295	20	47194750	G/A	0.5841	0.984819	0.9011 ± 0.040964	6.42E−07
rs6122700	20	47196231	T/C	0.6901	0.970105	0.8969 ± 0.043512	9.10E−07
rs6125386	20	47196367	A/G	0.6323	0.971916	0.8947 ± 0.041944	2.03E−07
rs6012487	20	47196557	G/C	0.6317	0.972012	0.8953 ± 0.041944	2.36E−07
rs4810826	20	47196647	T/C	0.6311	0.972755	0.8950 ± 0.041944	2.13E−07
rs4810827	20	47196809	T/C	0.6314	0.971384	0.8951 ± 0.041944	2.29E−07
rs6019297	20	47197230	C/T	0.625	0.958847	0.8893 ± 0.04214	4.71E−08

[a]Odds ratios were calculated with respect to A2

[b]Model 1 was adjusted for gender, number of years under observation over the age of 15, and first 10 principal components of genetic ancestry

[c]Model 2 was adjusted for gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry, diagnosis of any mental disorder as well as diagnosis of schizophrenia, bipolar disorders, affective disorders, autism spectrum disorders, anorexia, and ADHD

this gene has been highlighted in a GWAS for major depression conducted by the Psychiatric Genomics Consortium, which includes the iPSYCH dataset [52]. Both connections were supported by evidence from C–HiC interactions.

The lead SNP identified on chromosome 20 in Model 2, rs4809706, is also intergenic with the nearest gene being *PREX1* (distance = 47 kb), which has been associated with depression and autism-like behavior [53]. An intersection with GeneHancer database did not reveal any associations with known enhancers or promoters.

By querying the GWAS catalog and PubMed, we attempted to replicate 40 SNPs associated to suicidal ideation or suicide attempt in six previous GWAS studies [13, 14, 54–59], We were unable to replicate either of the genome-wide significant loci identified [13, 58]. Nevertheless, one of the suggestive associations on chromosome 11 [13], rs10437629 showed significance in both models 1 ($p = 0.029$) and 2 ($p = 0.03$) in the current data (Supplementary Table 3).

In the association study comparing individuals with major depressive disorders and suicide attempts to healthy

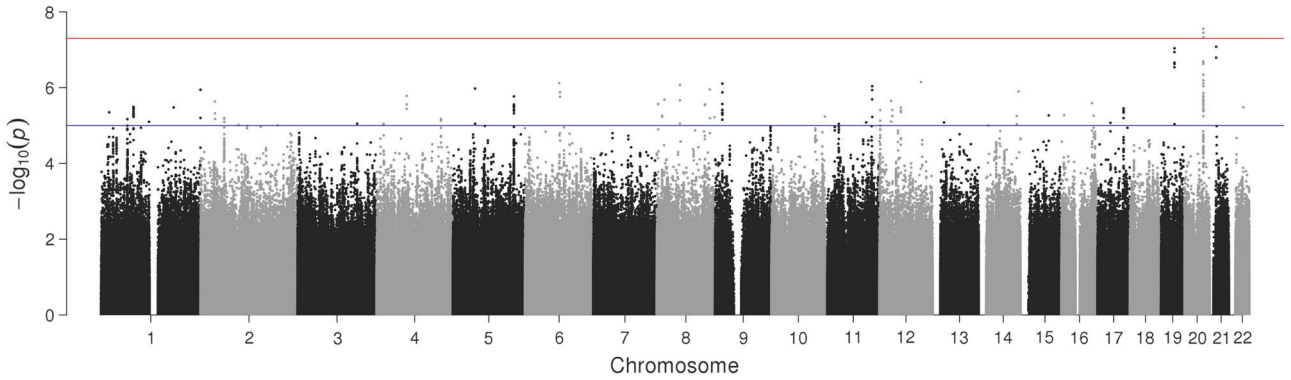


Fig. 2 Manhattan plot of genome-wide association study (GWAS) of suicide attempt adjusted for: gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry, diagnosis of any mental disorder as well as diagnosis of schizophrenia, bipolar disorders, affective disorders, autism spectrum disorders, anorexia, and ADHD (Model 2). *Note*: The red line marks the threshold of $p = 5 \times 10^{-8}$ for significant associations while the blue line marks the threshold of $p = 5 \times 10^{-6}$ for suggestive associations

controls, the lead SNP, rs4554696, is in the intronic region of the gene *PDE4B*, which has previously been implicated in schizophrenia and bipolar disorders [60, 61]. Studies in Japanese populations have associated *PDE4B* to panic disorders and major depression [62, 63]. A gene expression analysis from the GTEx portal (release v7) [64] revealed *PDE4B* being highly expressed in brain-related tissues (Supplementary Figure 2). An intersection with the GeneHancer database showed that four tag SNPs (LD: $r^2 > 0.8$ with the lead SNP) rs12062815 (p-value: 9.22×10^{-9}), rs12062901 (p-value: 7.8×10^{-9}), rs1318475 (p-value: 7.244×10^{-9}), and rs34482581 (p-value: 6.994×10^{-9}) fall within an enhancer for the gene *PDE4B* with the following brain tissue annotations: anterior caudate, cingulate gyrus, inferior temporal lobe, hippocampus middle (Enhancer ID: GH01H065961). The connection to the gene *PDE4B* was further validated by C–HiC interaction. Interestingly, this locus has not been observed to achieve genome-wide significance in previous studies of large-sample GWAS of major depressive disorder [52]; indicating that suicidal behavior could possibly be utilized as a secondary phenotype to explore severity within major depressive disorder. Behavioral and neurochemical characterization of *PDE4B* in an animal model revealed the *PDE4B*-knockout mice to have significantly reduced pre-pulse inhibition, decreased spontaneous locomotor activity, and enhanced response to amphetamine [65]. Additionally, *PDE4B*-deficient mice display anxiogenic-like behavior [66], further supporting the role of the *PDE4B* gene in the etiology of mental disorders.

Various causal pathways have been suggested [26]. While previous studies have identified loci associated with suicidal attempts in clinical or relatively small population samples [11], our findings add significant support to the role of genes in the etiology of suicidal behavior. Liability for mental disorders, such as affective disorders or borderline personality disorders, may be inherited and explain the link to suicidal behavior [20, 26, 27]. Secondly, the relationship between genes and suicidal behavior may be mediated by behavioral traits, such as impulsive and impulsive-aggressive behaviors [67, 68]. Furthermore, environmental factors, such as stressful life events and social constraints, might interact with genetic factors [27].

Heritability

A heritability of 4.6% is substantially lower than reported in the family and twin studies [9]. However, this type of studies cannot fully account for social contamination through social role models, social heritages, such as living conditions, and other shared environmental factors [69]. The GCTA estimates provide a lower bound estimate for genetic heritability, as not all genetic variation can be explained by genotyped SNPs [70]. Still, it is interesting that a heritability of 1.9% was maintained when adjusting for the presence of any and specific mental disorders; this estimate is also lower than the 17.4% noted in observational studies [8, 71]. The reduction from 4.6 to 1.9% indicates that parts of the genetic transmission might be explained through mental disorders.

Schizophrenia has been linked to an absolute suicide risk of 6.5% [19]. Nevertheless, a SNP heritability close to zero might be related to the relatively small number of individuals suffering from schizophrenia with a suicide attempt. Alternative explanations could be that genetic heterogeneity of suicide attempt is explained in part by diagnosed mental

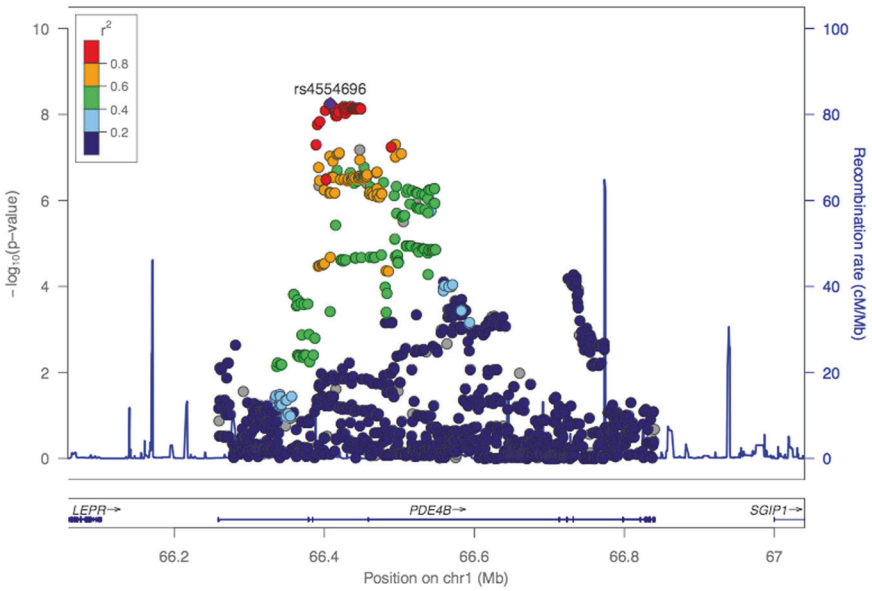


Fig. 3 Region plot of the association evidence in gene *PDE4B* in the chromosome 1p31.3 in a sub-group analysis of cases with affective disorders vs. healthy controls (hg19, LocusZoom viewer) [76].

disorders; or in conditions, such as schizophrenia, the genetic factors explaining suicide attempt are not distinct from those genetic factors increasing the risk for the disorder itself. Interestingly, a polygenic overlap between schizoaffective disorders and suicide attempt has previously been suggested [72]. The highest heritability was found for autism spectrum disorders; indicating that SNP heritability accounted for up to 10% of suicide attempts. This has not previously been documented and future studies might gain insights by assessing mental disorders separately.

Strengths and limitations

The Danish Newborn Screening Biobank presents an almost complete coverage and a validated approach for DNA extraction and genotyping was used [27, 35]. Furthermore, complete hospital records of psychiatric inpatients for the entire study period were available while data on emergency department and out-patients were included in 1995 when the oldest birth cohorts were 13 years of age [29], hence, keeping the probability of selection bias low. While previous studies had to rely on self-reported data [73], the phenotype was based on hospital records and measured consistently for all subjects. Using a large population-based sample with limited population heterogeneity adds additional strengths to this study. Furthermore, the inclusion of healthy individuals allowed for stratification by mental disorders [11].

Limitations should be acknowledged. About half of the genotyped SNPs were excluded due to the low-frequency exome SNPs in the iPSYCH array design [31]. Due to the sample acquisition, our findings pertain to suicide attempts among individuals below the age of 32 years; i.e., the age segment with the highest rates of suicide attempt [74]. Suicide attempt is under-recorded in Danish hospitals [74], and a substantial proportion of individuals do not seek hospital care after a suicide attempt [75]. By including a validated, wider proxy, any bias would likely render our estimates conservative; yet, the possibility of false positives cannot be excluded. For reasons of data availability, it was not feasible to examine different methods separately. Hospital-based diagnoses served as a proxy for severe mental disorders; while it is likely that the severe cases of mental disorders will be seen at the hospital, less severe cases might be missed. Also, we did not account for changes in the status of mental disorders over time. It would have been desirable to account for a range of different covariates, such as physical comorbidity and stressful life events, however, this was not feasible in this study.

Conclusion

This large, population-based GWAS of young adults identified SNPs with novel significant associations to suicide attempt while accounting for diagnoses of mental disorders. An association study comparing individuals with affective disorder and attempted suicide to healthy controls revealed a novel association in the intronic region of the gene PDE4B. Heritability estimates indicated that variation in a suicide attempt is associated with the genetic variation and that this association is, partially but perhaps not entirely, explained through mental disorders.

Acknowledgements The Lundbeck Foundation and the Novo Nordisk Foundation had no role in the design and conduct of the study;

Table 2 SNP heritability with respect to gender and mental disorders

	Suicide attempt *n* (%)[a]	Control group *n* (%)[a]	Model 1[b] h^2 [CI-95%]	*p*-Value	Model 2[c] h^2 [CI-95%]	*p*-Value
All	6024 (100.0)	44,240 (100.0)	4.6 [2.9–6.3]	<0.0001	1.9 [0.3–3.5]	0.008
Gender						
Males	1837 (30.5)	22,735 (51.4)	5.3 [2.4–8.2]	0.0001	2.8 [−0.1 to 5.7]	0.026
Females	4187 (69.5)	21,505 (48.6)	4.7 [1.9–7.5]	0.0003	1.9 [−0.7 to 4.5]	0.064
Mental disorders						
No diagnosis	132 (2.2)	13,294 (30.1)	3.3 [−1.0 to 7.6]	0.063		
Any diagnosis	5892 (97.8)	30,946 (70.0)	2.8 [0.7–4.9]	0.0027		
Diagnosis						
Schizophrenia	581 (9.6)	1579 (3.6)	0.0 [−27.5 to 27.5]	0.50		
Bipolar[d]	302 (5.0)	911 (2.1)	16.3 [−32.0 to 64.6]	0.254		
Affective disorders	4302 (71.4)	14,935 (33.8)	5.6 [1.9–9.3]	0.001		
Autism	402 (6.7)	7008 (15.8)	9.6 [1.1–18.1]	0.0115		
Anorexia	1254 (20.8)	7866 (17.8)	10.6 [−10.5 to 31.7]	0.148		
ADHD	350 (5.8)	2389 (5.4)	0.4 [−6.6 to 7.4]	0.45		

[a]Additional cases and controls were excluded from the GCTA analysis due to further adjustment for kinship by accounting for cryptic relatedness

[b]Model 1 was adjusted for gender, number of years under observation over the age of 15, and first 10 principal components of genetic ancestry. The h2 estimates are on the observed scale.

[c]Model 2 was adjusted for gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry, diagnosis of any mental disorder as well as diagnosis of schizophrenia, bipolar disorders, affective disorders, autism spectrum disorders, anorexia, and ADHD. The h2 estimates are on the observed scale.

[d]Bipolar disorders (ICD: F30–F31) was a subset of affective disorders (ICD: F30–F39)

collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Funding The study was supported by the Lundbeck Foundation (Grant numbers R102-A9118 and R155-2014-1724). Furthermore, this research has been conducted using the Danish National Biobank resource supported by the Novo Nordisk Foundation and the Lundbeck Foundation. Additionally, Dr. Thompson was supported by NIH 1R01GM104400.

Author contributions Appadurai, Dr. Wang, and Dr. Thompson had full access to all the data in the study and assume responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Erlangsen, Agerbo, Nordentoft, Thompson, Mors, Wang, Appadurai, Turecki, Mortensen, and Werge. Acquisition, analysis, or interpretation of data: All authors. Drafting of the manuscript: Erlangsen, Agerbo, Appadurai, Nordentoft, and Turecki. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: Appadurai, Wang, Thompson, Agerbo, Schork, Nudel, Erlangsen, and Turecki. Obtained funding: Nordentoft, Mortensen, Werge, Mors, and Hougaard. Administrative, technical, or material support: Thompson, Agerbo, Wang, Appadurai, Nordentoft, Mortensen, Werge, Mors, Hougaard, Bækvad-Hansen, and Bybjerg-Grauholm. Supervision: Agerbo, Thompson, Nordentoft, Turecki, and Mors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. Epidemiol Rev. 2008;30:133–54.
2. World Health Organization. Preventing suicide—a global imperative. Geneva: WHO; 2014.
3. Agerbo E, Nordentoft M, Mortensen PB. Familial, psychiatric, and socioeconomic risk factors for suicide in young people: nested case-control study. BMJ. 2002;325:74.
4. Runeson B, Asberg M. Family history of suicide among suicide victims. Am J Psychiatry. 2003;160:1525–6.
5. Tidemalm D, Runeson B, Waern M, Frisell T, Carlström E, Lichtenstein P, et al. Familial clustering of suicide risk: a total population study of 11.4 million individuals. Psychol Med. 2011; 41:2527–34.
6. Qin P, Agerbo E, Mortensen PB. Suicide risk in relation to socioeconomic, demographic, psychiatric and familial factors: a national register-based study of all suicides in Denmark, 1981–97. Am J Psychiatry. 2003;160:765–72.
7. Agerbo E, Qin P, Mortensen PB. Psychiatric illness, socioeconomic status, and marital status in people committing suicide: a matched case-sibling-control study. J Epidemiol Community Health. 2006;60:776–81.
8. Statham DJ, Heath AC, Madden PA, Bucholz KK, Bierut L, Dinwiddie SH, et al. Suicidal behaviour: an epidemiological and genetic study. Psychol Med. 1998;28:839–55.
9. Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet. 2016;387:1227–39.
10. Wender PH, Kety SS, Rosenthal D, Schulsinger F, Ortmann J, Lunde I. Psychiatric disorders in the biological and adoptive

families of adopted individuals with affective disorders. Arch Gen Psychiatry. 1986;43:923–9.
11. Sokolowski M, Wasserman J, Wasserman D. Genome-wide association studies of suicidal behaviors: a review. Eur Neuropsychopharmacol. 2014;24:1567–77.
12. Brezo J, Bureau A, Merette C, Jomphe V, Barker ED, Vitaro F, et al. Differences and similarities in the serotonergic diathesis for suicide attempts and mood disorders: a 22-year longitudinal gene-environment study. Mol Psychiatry. 2010;15:831–43.
13. Willour VL, Seifuddin F, Mahon PB, Jancic D, Pirooznia M, Steele J, et al. A genome-wide association study of attempted suicide. Mol Psychiatry. 2012;17:433–44.
14. Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, Sullivan PF, et al. Genome-wide association study of suicide attempts in mood disorder patients. Am J Psychiatry. 2010;167:1499–507.
15. Sokolowski M, Wasserman J, Wasserman D. An overview of the neurobiology of suicidal behaviors as one meta-system. Mol Psychiatry. 2015;20:56–71.
16. Baldessarini RJ, Hennen J. Genetics of suicide: an overview. Harv Rev Psychiatry. 2004;12:1–13.
17. Qin P. The impact of psychiatric illness on suicide: differences by diagnosis of disorders and by sex and age of subjects. J Psychiatr Res. 2011;45:1445–52.
18. Hawton K, van Heeringen K. Suicide. Lancet. 2009;373:1372–81.
19. Nordentoft M, Mortensen PB, Pedersen CB. Absolute risk of suicide after first hospital contact in mental disorder. Arch Gen Psychiatry. 2011;68:1058–64.
20. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. Am J Psychiatry. 2000;157:1552–62.
21. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013;381:1371–9.
22. Brent DA, Bridge J, Johnson BA, Connolly J. Suicidal behavior runs in families. A controlled family study of adolescent suicide victims. Arch Gen Psychiatry. 1996;53:1145–52.
23. McGirr A, Alda M, Seguin M, Cabot S, Lesage A, Turecki G. Familial aggregation of suicide explained by cluster B traits: a three-group family study of suicide controlling for major depressive disorder. Am J Psychiatry. 2009;166:1124–34.
24. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73:1–11.
25. Erlangsen A, Fedyszyn I. Danish nationwide registers for public health and health-related research. Scand J Public Health. 2015;43:333–9.
26. Pedersen CB. The Danish Civil Registration System. Scand J Public Health Suppl. 2011;39:22–5.
27. Norgaard-Pedersen B, Hougaard DM. Storage policies and use of the Danish Newborn Screening Biobank. J Inherit Metab Dis. 2007;30:530–6.
28. Nørgaard-Pedersen B, Simonsen H. Biological specimen banks in neonatal screening. Acta Paediatr Suppl. 1999;88:106–9.
29. Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. Scand J Public Health Suppl. 2011;39:54–7.
30. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health Suppl. 2011;39:30–3.
31. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, Grove J, Agerbo E, Baekvad-Hansen M, et al. The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol Psychiatry*. 2017;23:6–14.
32. International statistical classification of diseases and related health problems, 10th revision. Geneva: World Health Organization; 2007. Available from: http://apps.who.int/classifications/apps/icd/icd10online/ [updated 11/24/2009].
33. Laursen TM, Trabjerg BB, Mors O, Borglum AD, Hougaard DM, Mattheisen M, et al. Association of the polygenic risk score for schizophrenia with mortality and suicidal behavior—a Danish population-based study. Schizophr Res. 2017;184:122–7.
34. Pedersen MG, Mortensen P, Norgaard-Pedersen B, Postolache TT. *Toxoplasma gondii* infection and self-directed violence in mothers. Arch Gen Psychiatry. 2012;69:1123–30.
35. Hollegaard MV, Grauholm J, Borglum A, Nyegaard M, Norgaard-Pedersen B, Orntoft T, et al. Genome-wide scans using archived neonatal dried blood spot samples. BMC Genomics. 2009;10:297.
36. Hollegaard MV, Sørensen KM, Petersen HK, Arnardottir MB, Nørgaard-Pedersen B, Thorsen P, et al. Whole genome amplification and genetic analysis after extraction of proteins from dried blood spots. Clin Chem. 2007;53:1161–2.
37. Agerbo E, Mortensen PB, Wiuf C, Pedersen MS, McGrath J, Hollegaard MV, et al. Modelling the contribution of family history and variation in single nucleotide polymorphisms to risk of schizophrenia: a Danish national birth cohort-based study. Schizophr Res. 2012;134:246–52.
38. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. Nat Genet. 2016;48:817–20.
39. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
40. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
41. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.
42. Schork AJ, Won H, Appadurai V, Nudel R, Gandal M, Delaneau O, et al. A genome-wide association study for shared risk across major psychiatric disorders in a nation-wide birth cohort implicates fetal neurodevelopment as a key mediator. *bioRxivorg*. 2017.
43. Chang CC, Chow C, Tellier L, Vattikuti S, Purcell S, Lee J. Second generation PLINK: rising to the challenge of larger and richer datasets. GigaScience; 2015;4:7.
44. Turner S. qqman: an R package for visualizing GWAS results using Q–Q and Manhattan plots.
45. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
46. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). 2017;2017:bax028.
47. UCSC Genome Browser Database. Santa Cruz, USA: University of California. Available from: https://genome.ucsc.edu/cgi-bin/hgLiftOver [cited 2018].
48. Quinlan AR. BEDTools: the Swiss-Army tool for genome feature analysis. Curr Protoc Bioinform. 2002. https://doi.org/10.1002/0471250953.bi1112s47.
49. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580.
50. Gupta S, Racaniello AA. Neuroleptic malignant syndrome associated with amoxapine and lithium in an older adult. Ann Clin Psychiatry. 2000;12:107–9.
51. Tesfaye H, Prusa R, Doupovcova J. [Hypokalaemia in a suicide attempt of an adolescent girl]. Casopís Lékařů Českých. 2008;147:333–6.

52. Wray NR, Sullivan PF. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–81.
53. Li J, Chai A, Wang L, Ma Y, Wu Z, Yu H, et al. Synaptic P-Rex1 signaling regulates hippocampal long-term depression and autism-like social behavior. Proc Natl Acad Sci USA. 2015;112:E6964–72.
54. Mullins N, Perroud N, Uher R, Butler AW, Cohen-Woods S, Rivera M, et al. Genetic relationships between suicide attempts, suicidal ideation and major psychiatric disorders: a genome-wide association and polygenic scoring study. Am J Med Genet B Neuropsychiatr Genet. 2014;165:428–37.
55. Galfalvy H, Haghighi F, Hodgkinson C, Goldman D, Oquendo MA, Burke A, et al. A genome-wide association study of suicidal behavior. Am J Med Genet B Neuropsychiatr Genet. 2015;168:557–63.
56. Zai CC, Gonçalves VF, Tiwari AK, Gagliano SA, Hosang G, De Luca V, et al. A genome-wide association study of suicide severity scores in bipolar disorder. J Psychiatr Res. 2015;65:23–9.
57. Perroud N, Uher R, Ng M, Guipponi M, Hauser J, Henigsberg N, et al. Genome-wide association study of increasing suicidal ideation during antidepressant treatment in the GENDEP project. Pharm J. 2012;12:68.
58. Stein MB, Ware EB, Mitchell C, Chen CY, Borja S, Cai T, et al. Genome wide association studies of suicide attempts in US soldiers. Am J Med Genet B Neuropsychiatr Genet. 2017;174:786–97.
59. GWAS Catalog. Bethesda, MD, USA: National Human Genome Research Institute. Available from: https://www.ebi.ac.uk/gwas/home [cited 2018].
60. Feng Y, Cheng D, Zhang C, Li Y, Zhang Z, Wang J, et al. Association of PDE4B polymorphisms with susceptibility to schizophrenia: a meta-analysis of case-control studies. PLoS *ONE*. 2016;11:e0147092.
61. McDonald ML, MacMullen C, Liu DJ, Leal SM, Davis RL. Genetic association of cyclic AMP signaling genes with bipolar disorder. Transl Psychiatry. 2012;2:e169.
62. Otowa T, Kawamura Y, Sugaya N, Yoshida E, Shimada T, Liu X, et al. Association study of PDE4B with panic disorder in the Japanese population. Prog Neuropsychopharmacol Biol Psychiatry. 2011;35:545–9.
63. Iga J, Ueno S, Yamauchi K, Numata S, Tayoshi-Shibuya S, Kinouchi S, et al. The Val66Met polymorphism of the brain-derived neurotrophic factor gene is associated with psychotic feature and suicidal behavior in Japanese major depressive patients. Am J Med Genet B Neuropsychiatr Genet. 2007;144B:1003–6.
64. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
65. Siuciak JA, McCarthy SA, Chapin DS, Martin AN. Behavioral and neurochemical characterization of mice deficient in the phosphodiesterase-4B (PDE4B) enzyme. Psychopharmacology (Berl). 2008;197:115–26.
66. Zhang H-T, Huang Y, Masood A, Stolinski LR, Li Y, Zhang L, et al. Anxiogenic-like behavioral phenotype of mice deficient in phosphodiesterase 4B (PDE4B). Neuropsychopharmacology. 2007;33:1611.
67. Turecki G. The molecular bases of the suicidal brain. Nat Rev Neurosci. 2014;15:802–16.
68. Turecki G. Suicidal behavior: is there a genetic predisposition? Bipolar Disord. 2001;3:335–49.
69. Brent DA, Mann JJ. Family genetic studies, suicide, and suicidal behavior. Am J Med Genet C Semin Med Genet. 2005;133C:13–24.
70. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
71. Mann JJ, Arango VA, Avenevoli S, Brent DA, Champagne FA, Clayton P, et al. Candidate endophenotypes for genetic studies of suicidal behavior. Biol Psychiatry. 2009;65:556–63.
72. Sokolowski M, Wasserman J, Wasserman D. Polygenic associations of neurodevelopmental genes in suicide attempt. Mol Psychiatry. 2015;21:1381.
73. Gross JA, Bureau A, Croteau J, Galfalvy H, Oquendo MA, Haghighi F, et al. A genome-wide copy number variant study of suicidal behavior. PLoS *ONE*. 2015;10:e0128369.
74. Morthorst B, Soegaard B, Nordentoft M, Erlangsen A. Incidence rates of deliberate self-harm in Denmark 1994–2011. Crisis. 2016;37:256–64.
75. Hawton K, Saunders KE, O'Connor RC. Self-harm and suicide in adolescents. Lancet. 2012;379:2373–82.
76. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336–7.

Affiliations

Annette Erlangsen[1,2,3,4] · Vivek Appadurai[1,5] · Yunpeng Wang[1,6,7] · Gustavo Turecki[8] · Ole Mors[1,9] · Thomas Werge[1,5,10] · Preben B. Mortensen[1,11] · Anna Starnawska[1,12] · Anders D. Børglum[1,12,13] · Andrew Schork[1,5] · Ron Nudel[1,5] · Marie Bækvad-Hansen[1,14] · Jonas Bybjerg-Grauholm[1,14] · David M. Hougaard[1,14] · Wesley K. Thompson[1,6,7,9,15] · Merete Nordentoft[1,2,10,16] · Esben Agerbo[1,11]

[1] The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark

[2] Danish Research Institute for Suicide Prevention, Mental Health Centre Copenhagen, Copenhagen, Denmark

[3] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[4] Center of Mental Health Research, Australian National University, Canberra, Australia

[5] Institute of Biological Psychiatry, Mental Health Center St. Hans, Mental Health Services Copenhagen, Roskilde, Denmark

[6] Norwegian Centre for Mental Disorders Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[7] Division of Mental Health and Addiction, University of Oslo, Oslo, Norway

[8] McGill Group for Suicide Studies, Douglas Hospital Research Centre, Department of Psychiatry, McGill University, Montreal, Canada

[9] Psychosis Research Unit, Aarhus University Hospital,

Risskov, Denmark

[10] Institute of Clinical Medicine, Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark

[11] National Centre for Register-based Research (NCRR) and Centre for Integrated Register-based Research (CIRRAU), Aarhus University, Aarhus, Denmark

[12] Department of Biomedicine and Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark

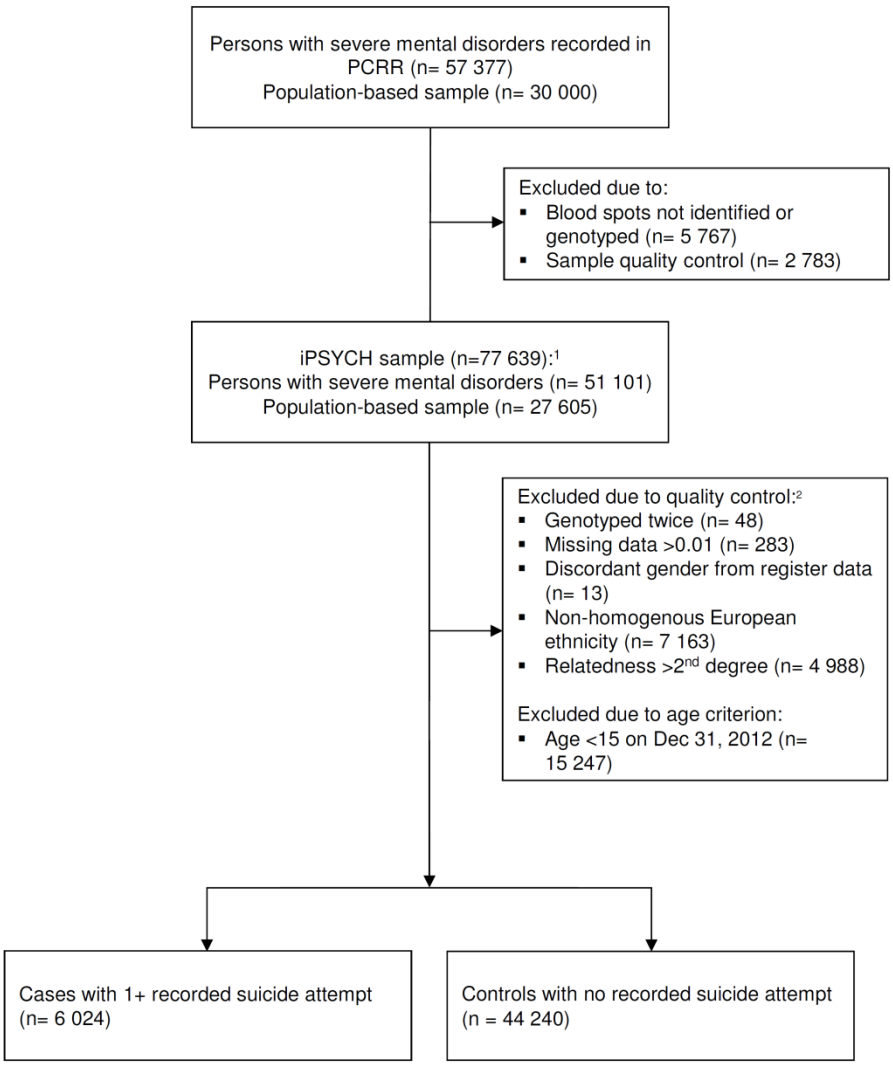
[13] Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Denmark

[14] Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark

[15] Division of Biostatistics, Department of Family Medicine and Public Health, University of California, San Diego, CA, USA

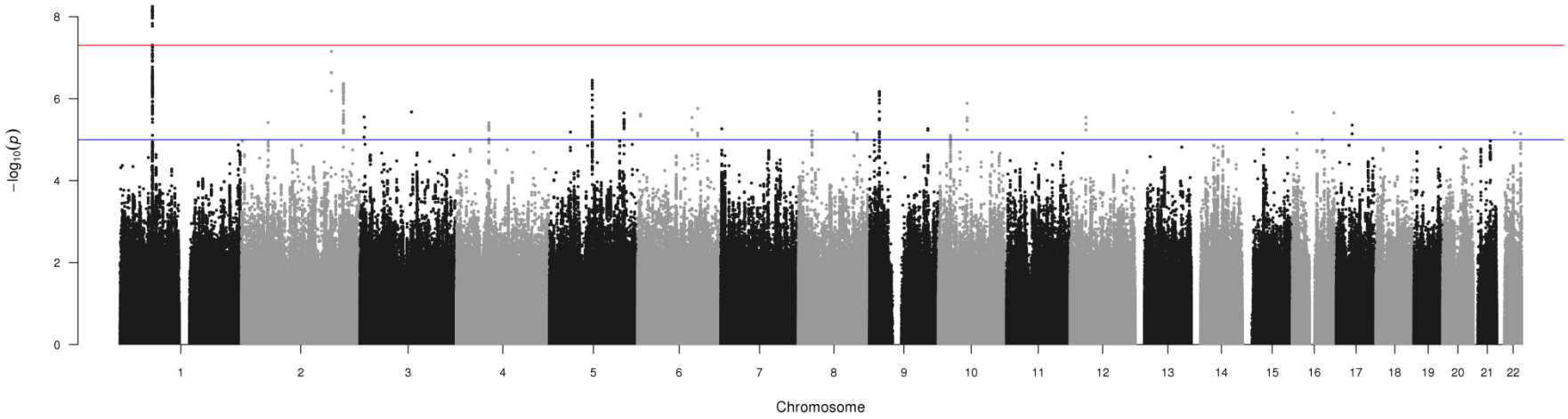
[16] Research Unit, Mental Health Centre Copenhagen, University of Copenhagen, Copenhagen, Denmark

Supplementary Figure 1. Flow diagram.



Notes:
Abbreviation: PCRR: Psychiatric Central Research Register.
[1] The two sub/groups, persons with severe mental disorders and the population-based sample, were not mutually exclusive, i.e. some members of the population-based sample had been recorded with severe mental disorders.
[2] The listed reasons for exclusion were not mutually exclusive.

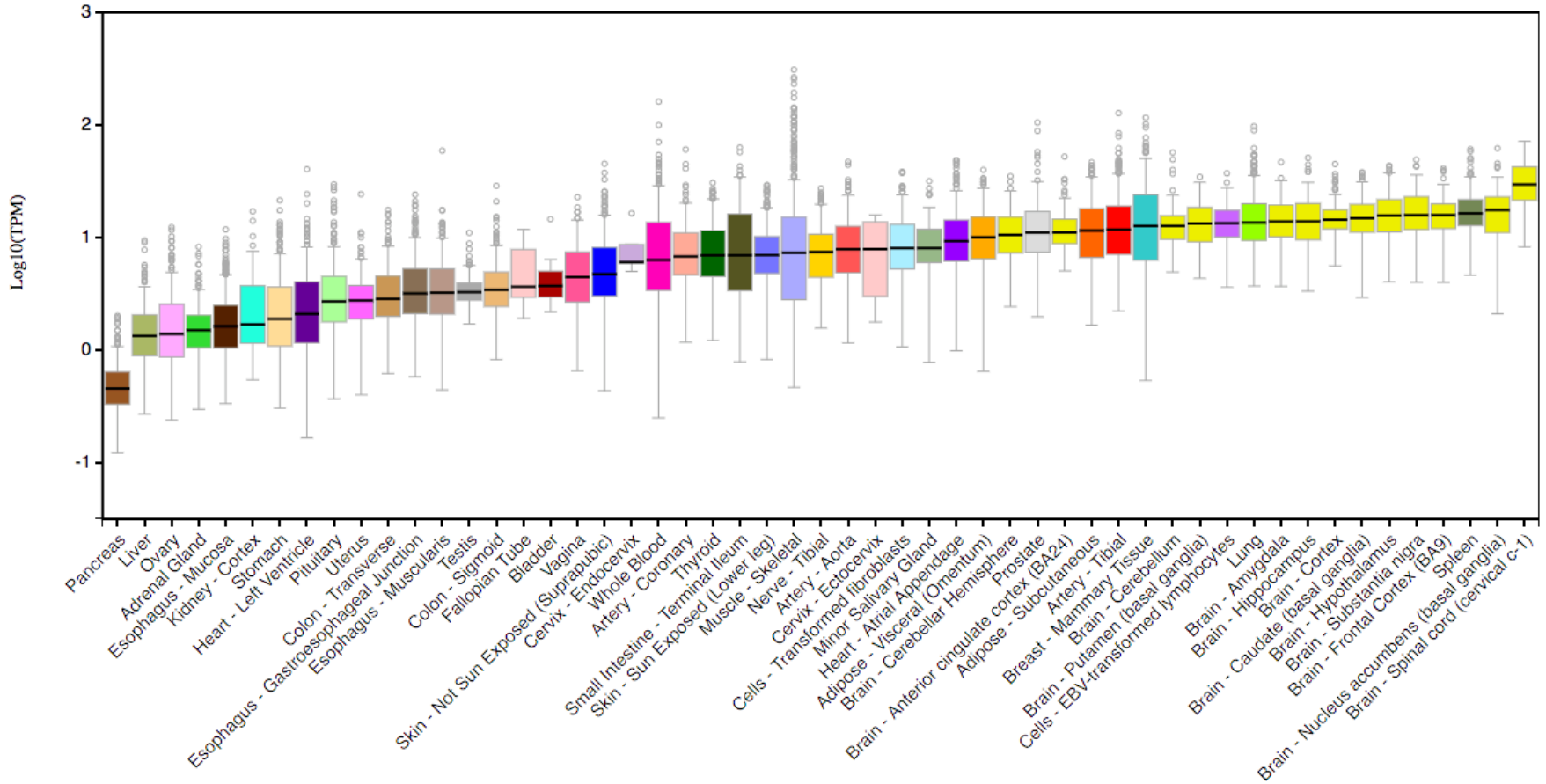
Supplementary Figure 2. GWAS of sub-group analysis of cases with affective disorders vs. healthy controls.[1]



[1] The model was adjusted for socio-demographic covariates: gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry.

Note: The red line marks the threshold of $p=5 \times 10^{-8}$ for significant associations while the blue line marks the threshold of $p=5 \times 10^{-6}$ for suggestive associations.

Supplementary Figure 3. *PDE4B* Gene Expression in GTEx Tissues.



Supplementary Table 1. Sample characteristics.

	Suicidal behavior (N=6,024)	**No suicidal behavior** (N=44,240)	χ^2 (df); p-value
Age (mean, years ± SD)	24.3 ± 4.11	22.9 ±4.6	
Age (%)			
15-19	1 079 (17.9%)	14 294 (32.3%)	537.3 (3); p<0.0001
20-24	2 239 (37.2%)	14 489 (32.8%)	
25-29	2 143 (35.6%)	12 162 (27.5%)	
30-34	564 (9.4%)	3 296 (7.5%)	
Gender (% male)	1,837 (30.5%)	22 735 (51.4%)	926.9 (1); p<0.0001
Birth year (%)			
1981-1984	1 328 (22.1%)	7 733 (17.5%)	595.9 (3); p<0.0001
1985-1989	2 342 (38.9%)	13 250 (30.0%)	
1990-1994	1 904 (31.6%)	14 931 (33.8%)	
1995-1999	450 (7.5%)	8 326 (18.8%)	
Mental disorders (%)			
None	132 (2.2%)	13 294 (30.1%)	2 102.0 (1); p<0.0001
Any	5 892 (97.8%)	30 946 (70.0%)	
Schizophrenia	581 (9.6%)	1 579 (3.6%)	475.0 (1); p<0.0001
Bipolar disorders	302 (5.0%)	911 (2.1%)	195.9 (1); p<0.0001
Affective disorders	4 302 (71.4%)	14 935 (33.8%)	3 181.3 (1); p<0.0001
Autism spectrum	402 (6.7%)	7 008 (15.8%)	354.7 (1); p<0.0001
ADHD	1 254 (20.8%)	7 866 (17.8%)	32.8 (1); p<0.0001
Anorexia	350 (5.8%)	2 389 (5.4%)	1.7 (1); p=0.188
Number of suicide attempts (%)			
1	3 294 (54.7%)		
2	1 352 (22.4%)		
3-4	889 (14.8%)		
5-9	369 (6.1%)		
10+	120 (2.0%)		

Supplementary Table 2. SNP loci showing significant or suggestive association with suicide attempt among patients with affective disorders in comparison to healthy controls in adjusted GWAS.

SNP	Chromosome	Position	A1/A2	Minor allele frequency	INFO	Odds ratio [CI-95%]	P-value
rs4655809	1	66389097	C / T	0.358	0.961614	0.8587+/-0.054684	5.02E-08
rs4655591	1	66390808	A / T	0.3604	0.97113	0.8550+/-0.054488	1.70E-08
rs2310752	1	66392405	G / A	0.4127	0.969206	0.8680+/-0.053116	1.71E-07
rs60745892	1	66393106	T / TTA	0.3355	0.971848	0.8671+/-0.055468	4.55E-07
rs2997084	1	66393732	C / G	0.4146	0.972114	0.8714+/-0.05292	3.43E-07
rs10889589	1	66394025	C / G	0.3605	0.973138	0.8545+/-0.054292	1.46E-08
rs4255357	1	66399895	G / A	0.5594	0.975406	1.1425+/-0.052136	5.75E-07
rs11585651	1	66401440	T / C	0.361	0.977899	0.8524+/-0.054292	8.10E-09
rs12058296	1	66402424	C / A	0.3362	0.979404	0.8660+/-0.055272	3.27E-07
rs2310754	1	66403092	G / A	0.5592	0.980727	1.1457+/-0.052136	3.10E-07
rs5774777	1	66406683	C / CT	0.2622	0.976222	0.8592+/-0.059976	6.82E-07
rs4503327	1	66406807	T / C	0.3613	0.982543	0.8513+/-0.054096	5.75E-09
rs7528604	1	66407352	G / A	0.4147	0.981715	0.8662+/-0.052724	9.37E-08
rs7550837	1	66407396	A / G	0.2622	0.976417	0.8590+/-0.059976	6.59E-07
rs12088813	1	66407700	A / C	0.2622	0.976493	0.8590+/-0.059976	6.57E-07
rs4554696	1	66408011	C / T	0.3614	0.982875	0.8512+/-0.054096	5.63E-09
rs12730848	1	66408646	C / T	0.2622	0.976692	0.8591+/-0.05978	6.70E-07
rs11208774	1	66410109	C / T	0.5595	0.982289	1.1462+/-0.052136	2.80E-07
rs11208775	1	66410541	C / T	0.5591	0.983137	1.1458+/-0.05194	2.95E-07
rs3009872	1	66411400	T / C	0.4152	0.983866	0.8675+/-0.052724	1.22E-07
rs11208776	1	66411733	A / T	0.5591	0.98366	1.1459+/-0.05194	2.88E-07
rs7527901	1	66412760	C / T	0.3611	0.984441	0.8526+/-0.054096	7.82E-09
rs7531270	1	66413731	G / A	0.3605	0.986969	0.8524+/-0.054096	7.33E-09
rs7543245	1	66413965	T / A	0.2627	0.979622	0.8594+/-0.05978	6.67E-07
rs6690101	1	66414039	T / C	0.5607	0.988771	1.1458+/-0.05194	2.77E-07
rs6680334	1	66414440	G / T	0.3605	0.988398	0.8524+/-0.054096	7.18E-09

rs35130999	1	66415187	A / T	0.3611	0.991349	0.8529+/-0.054096	7.65E-09
rs10889590	1	66415481	G / A	0.3599	0.993421	0.8543+/-0.0539	1.06E-08
rs1937458	1	66416309	C / T	0.3603	0.995445	0.8545+/-0.0539	1.08E-08
rs1937456	1	66416747	C / T	0.3603	0.996158	0.8545+/-0.0539	1.06E-08
rs1937455	1	66416939	A / G	0.4164	1	0.8670+/-0.052332	8.57E-08
rs11208779	1	66417145	G / C	0.5578	0.997177	1.1469+/-0.051744	1.99E-07
rs1937454	1	66417187	G / C	0.3604	0.996439	0.8545+/-0.0539	1.04E-08
rs7547416	1	66419087	G / A	0.5839	0.997819	1.1536+/-0.052332	8.38E-08
rs2997088	1	66419142	C / A	0.6397	0.99686	1.1712+/-0.0539	8.89E-09
rs6421482	1	66419905	A / G	0.5843	0.996134	1.1542+/-0.052332	7.86E-08
rs17417915	1	66421606	A / C	0.2837	0.995074	0.8598+/-0.058016	3.24E-07
rs10789205	1	66426100	T / C	0.6399	1	1.1729+/-0.0539	6.45E-09
rs12760902	1	66426251	C / T	0.2837	0.99524	0.8598+/-0.058016	3.19E-07
rs4638095	1	66426350	A / G	0.6398	0.999633	1.1722+/-0.0539	7.28E-09
rs12062815	1	66428569	A / G	0.3596	0.998499	0.8540+/-0.0539	9.22E-09
rs12062901	1	66428788	A / C	0.3601	0.999341	0.8533+/-0.0539	7.80E-09
rs1318475	1	66430192	G / C	0.6398	0.999583	1.1722+/-0.0539	7.24E-09
rs34482581	1	66430577	TTATACCCAGTGGA / T	0.6398	0.999567	1.1724+/-0.0539	6.99E-09
rs12562632	1	66431017	T / A	0.2838	0.995806	0.8602+/-0.05782	3.43E-07
rs140361678	1	66432175	A / C	0.2837	0.995125	0.8597+/-0.058016	3.16E-07
rs7513141	1	66432913	G / A	0.6389	0.997395	1.1727+/-0.0539	6.81E-09
rs7519259	1	66434743	G / A	0.5586	0.995135	1.1464+/-0.051744	2.26E-07
rs2186123	1	66435842	G / A	0.6401	0.997858	1.1730+/-0.0539	6.46E-09
rs34884275	1	66435900	CT / C	0.3675	0.97397	0.8518+/-0.054292	7.03E-09
rs6691929	1	66435953	A / G	0.2835	0.995388	0.8598+/-0.058016	3.20E-07
rs1937441	1	66437733	A / G	0.2832	0.996778	0.8596+/-0.05782	3.04E-07
rs1937440	1	66437771	A / G	0.6401	0.997458	1.1723+/-0.0539	7.45E-09
rs2503207	1	66438655	G / A	0.6401	0.997512	1.1725+/-0.0539	7.18E-09
rs1937439	1	66439127	A / C	0.283	0.996195	0.8592+/-0.058016	2.82E-07
rs1937438	1	66439291	C / T	0.6401	0.997458	1.1723+/-0.0539	7.45E-09
rs1937437	1	66439923	T / C	0.6401	0.997458	1.1723+/-0.0539	7.45E-09
rs2310819	1	66440096	C / T	0.5593	0.993559	1.1435+/-0.051744	3.91E-07

rs1937436	1	66441329	G / A	0.2832	0.996805	0.8596+/-0.05782	3.04E-07
rs1937434	1	66441795	G / T	0.6401	0.997525	1.1725+/-0.0539	7.17E-09
rs2503201	1	66442025	T / C	0.6401	0.997469	1.1723+/-0.0539	7.44E-09
rs10889591	1	66442338	C / T	0.6401	0.997465	1.1723+/-0.0539	7.45E-09
rs2503199	1	66444017	C / T	0.6401	0.997452	1.1723+/-0.0539	7.46E-09
rs67038944	1	66444113	CAGAA / C	0.2832	0.996766	0.8596+/-0.05782	0.000000306
rs1937433	1	66444183	C / G	0.559	0.994753	1.1440+/-0.051744	3.52E-07
rs66752967	1	66444221	ATGAT / A	0.2832	0.996852	0.8596+/-0.05782	3.04E-07
rs2503197	1	66444649	G / A	0.6402	0.997558	1.1723+/-0.0539	7.42E-09
rs12564578	1	66445516	G / A	0.2831	0.997667	0.8596+/-0.05782	3.03E-07
rs6684069	1	66445858	G / A	0.6402	0.99786	1.1726+/-0.0539	7.10E-09
rs1937453	1	66446672	T / A	0.7168	0.998132	1.1639+/-0.05782	2.72E-07
rs1937457	1	66446787	C / T	0.2831	0.997689	0.8596+/-0.05782	3.04E-07
rs35405572	1	66447228	CT / C	0.7214	0.95167	1.1784+/-0.059584	6.65E-08
rs6690398	1	66447394	G / A	0.5855	0.995114	1.1523+/-0.052332	1.15E-07
rs2503196	1	66447523	T / G	0.6402	0.997878	1.1726+/-0.0539	7.10E-09
rs2096406	1	66448005	C / T	0.7168	0.998159	1.1639+/-0.05782	2.72E-07
rs2489918	1	66448695	C / T	0.6404	0.996956	1.1724+/-0.0539	7.47E-09
rs2503194	1	66449215	G / A	0.7171	0.997137	1.1635+/-0.05782	2.96E-07
rs34844953	1	66449460	C / CA	0.2831	0.997706	0.8597+/-0.05782	3.04E-07
rs35528758	1	66450051	T / C	0.2831	0.997848	0.8595+/-0.05782	2.93E-07
rs57323522	1	66450601	ACT / A	0.2829	0.997175	0.8592+/-0.05782	2.81E-07
rs1317611	1	66451777	C / G	0.2831	0.99789	0.8595+/-0.05782	2.93E-07
rs12749570	1	66452494	C / T	0.2831	0.997904	0.8595+/-0.05782	2.93E-07
rs2186120	1	66453163	G / A	0.5577	0.992841	1.1482+/-0.051744	1.69E-07
rs59255680	1	66453384	T / C	0.2831	0.997947	0.8595+/-0.05782	2.93E-07
rs5023204	1	66453835	T / C	0.283	0.997909	0.8597+/-0.05782	3.03E-07
rs4531269	1	66455345	G / A	0.2829	0.997469	0.8593+/-0.05782	2.81E-07
rs2186118	1	66456465	C / A	0.2829	0.997602	0.8593+/-0.05782	2.81E-07
rs149108375	1	66456765	ATT / A	0.7091	0.977963	1.1615+/-0.058016	4.05E-07
rs2253745	1	66457244	A / G	0.7171	1	1.1643+/-0.05782	2.52E-07
rs12566724	1	66460427	T / C	0.2616	0.991472	0.8601+/-0.059584	6.84E-07
rs12567613	1	66460513	A / G	0.2616	0.991186	0.8602+/-0.059584	7.07E-07

rs6664196	1	66460585	T / G	0.2616	0.991147	0.8602+/-0.059584	6.99E-07
rs2503187	1	66461117	C / T	0.7383	0.990965	1.1632+/-0.059584	6.31E-07
rs2503185	1	66461401	A / G	0.4409	0.995889	0.8756+/-0.051744	4.85E-07
rs1937444	1	66463920	A / C	0.7394	0.989919	1.1633+/-0.059584	6.77E-07
rs10493393	1	66465678	A / G	0.2603	0.989693	0.8602+/-0.05978	7.73E-07
rs1937443	1	66469643	C / G	0.5853	0.99663	1.1482+/-0.052332	2.28E-07
rs2186122	1	66470206	A / T	0.5853	0.996728	1.1484+/-0.052332	2.19E-07
rs7534143	1	66470379	T / G	0.5632	1	1.1415+/-0.051744	5.35E-07
rs7522876	1	66470915	G / T	0.2604	0.989419	0.8606+/-0.05978	8.29E-07
rs2503182	1	66471099	A / G	0.7396	0.989328	1.1631+/-0.05978	7.02E-07
rs2489907	1	66473921	T / C	0.7391	0.985545	1.1621+/-0.05978	8.37E-07
rs61796569	1	66476437	C / T	0.2607	0.982576	0.8592+/-0.059976	6.81E-07
rs6588177	1	66476754	A / T	0.7369	0.980075	1.1634+/-0.05978	6.99E-07
rs1937450	1	66478840	T / G	0.5646	0.974957	1.1456+/-0.052528	3.83E-07
rs2455021	1	66489304	C / G	0.6591	0.973276	1.1654+/-0.055272	5.65E-08
rs4360504	1	66493152	T / C	0.5619	0.968635	1.1415+/-0.052528	7.69E-07
rs9659943	1	66493534	C / A	0.341	0.972495	0.8578+/-0.055076	5.08E-08
rs7525758	1	66494755	A / G	0.3426	0.98541	0.8586+/-0.05488	4.97E-08
rs12734198	1	66495075	T / A	0.3362	0.990148	0.8610+/-0.055076	9.63E-08
rs2455032	1	66496139	T / G	0.6347	0.991198	1.1479+/-0.053704	4.71E-07
rs2503221	1	66502677	T / C	0.6637	0.990425	1.1623+/-0.055076	8.18E-08
exm2252665	1	66509014	A / C	0.5676	1	1.1397+/-0.05194	7.65E-07
rs11208787	1	66509772	G / A	0.5677	0.999743	1.1407+/-0.05194	6.47E-07
rs1354060	1	66511404	A / G	0.567	0.998271	1.1418+/-0.05194	5.48E-07
rs1354059	1	66517279	A / T	0.5675	0.999568	1.1412+/-0.05194	5.95E-07
rs1500950	1	66522271	G / A	0.5676	0.99911	1.1408+/-0.05194	6.51E-07
rs11208790	1	66522409	G / C	0.5676	0.999107	1.1408+/-0.05194	6.53E-07
rs2202061	1	66523826	A / G	0.5676	0.999102	1.1406+/-0.05194	6.73E-07
rs58128575	1	66524024	A / T	0.5676	0.999107	1.1406+/-0.05194	6.76E-07
rs6588183	1	66524258	C / A	0.5676	0.999111	1.1405+/-0.05194	6.80E-07
rs1500961	1	66530369	G / A	0.5677	0.998728	1.1403+/-0.05194	7.17E-07
rs1392821	1	66533755	A / T	0.5678	0.998406	1.1399+/-0.05194	7.67E-07
rs1392820	1	66533859	G / A	0.5677	0.998509	1.1401+/-0.05194	7.37E-07

rs1500959	1	66538030	G / A	0.5678	0.998397	1.1389+/-0.05194	9.10E-07
rs72502638	1	66538698	CT / C	0.5454	0.957361	1.1425+/-0.05292	8.02E-07
rs1392817	1	66539456	A / G	0.535	0.979421	1.1424+/-0.052136	5.70E-07
rs868321	1	66541838	A / G	0.5349	0.978392	1.1426+/-0.052136	5.62E-07
rs7539350	1	66546376	G / A	0.5346	0.975843	1.1432+/-0.052332	5.31E-07
rs6661750	1	66546884	A / G	0.5346	0.975723	1.1432+/-0.052332	5.30E-07
rs62197059	2	184606702	A / C	0.0168	0.709053	1.7766+/-0.217756	2.32E-07
rs62198260	2	184606704	C / T	0.0168	0.709472	1.7771+/-0.217756	2.30E-07
rs182087934	2	184607434	G / A	0.0165	0.706854	1.8276+/-0.219324	7.04E-08
rs62199592	2	184723500	A / G	0.021	0.781197	1.6071+/-0.186788	6.49E-07
rs35293936	2	209227348	C / T	0.1081	0.999121	1.2263+/-0.079772	5.34E-07
rs12991283	2	209227547	A / G	0.1081	0.999028	1.2263+/-0.079772	5.35E-07
rs12998277	2	209228206	G / T	0.1081	0.999227	1.2262+/-0.079772	5.38E-07
rs28885800	2	209228648	A / G	0.1081	0.999165	1.2262+/-0.079772	5.41E-07
rs112057671	2	209229587	AT / A	0.1081	0.999287	1.2261+/-0.079772	5.45E-07
rs13395035	2	209230011	C / T	0.1081	0.999463	1.2260+/-0.079772	5.48E-07
rs67660353	2	209230737	G / A	0.1081	0.999436	1.2260+/-0.079772	5.51E-07
rs11891163	2	209233004	T / C	0.1081	0.999768	1.2257+/-0.079772	5.64E-07
rs13400356	2	209233095	T / C	0.1081	0.999818	1.2268+/-0.079772	5.04E-07
rs35217101	2	209233517	C / T	0.1081	1	1.2257+/-0.079772	5.67E-07
rs374532399	2	209242834	G / A	0.0838	0.98675	1.2530+/-0.089964	8.73E-07
rs16841143	2	209249574	G / A	0.1066	0.990594	1.2309+/-0.080556	4.27E-07
rs67757182	2	209251446	A / G	0.084	0.986487	1.2562+/-0.089964	6.53E-07
rs58081564	2	209252857	AC / A	0.1098	0.96976	1.2267+/-0.08036	6.35E-07
rs10206991	2	209255020	G / A	0.1065	0.990467	1.2309+/-0.080556	4.32E-07
rs67393765	2	209259217	A / C	0.0839	0.986468	1.2563+/-0.089964	6.59E-07
rs147754876	2	209260569	T / C	0.0839	0.986471	1.2563+/-0.089964	6.59E-07
rs112603275	2	209262300	T / C	0.0839	0.986438	1.2563+/-0.089964	6.60E-07
rs11897777	2	209264186	A / G	0.1064	0.990603	1.2310+/-0.080556	4.33E-07
rs13397965	2	209266375	G / A	0.1063	0.989101	1.2308+/-0.080556	4.47E-07
rs7608913	2	209272595	A / G	0.0839	0.985414	1.2560+/-0.089964	6.80E-07
rs376161923	2	209273259	GA / G	0.0875	0.956569	1.2501+/-0.089376	9.62E-07
rs3886855	2	209273632	T / C	0.1063	0.988696	1.2311+/-0.080752	4.36E-07

rs11904715	2	209274275	A / G	0.0843	0.983488	1.2546+/-0.089964	7.57E-07
rs72988764	2	209274874	G / A	0.0841	0.983997	1.2585+/-0.089964	5.34E-07
rs13414881	2	209276304	G / A	0.1063	0.990322	1.2306+/-0.080556	4.53E-07
rs67920705	2	209277115	A / C	0.0838	0.986706	1.2563+/-0.089964	6.66E-07
rs11902124	2	209280931	T / C	0.0837	0.986868	1.2563+/-0.089964	6.69E-07
rs11902410	2	209281515	T / C	0.0837	0.986913	1.2564+/-0.089964	6.69E-07
rs61193569	2	209292374	A / G	0.0834	0.98876	1.2561+/-0.09016	6.99E-07
rs13000248	2	209294818	T / A	0.1059	0.991971	1.2306+/-0.080752	4.71E-07
rs1347664	2	209295729	C / T	0.0832	0.991081	1.2558+/-0.09016	7.28E-07
rs1816612	2	209296070	A / G	0.0832	0.991171	1.2557+/-0.09016	7.30E-07
rs10497901	2	209303908	A / G	0.0831	0.991548	1.2557+/-0.09016	7.34E-07
rs1437413	2	209306793	G / C	0.0831	0.991458	1.2556+/-0.09016	7.39E-07
rs16841221	2	209312887	C / T	0.0831	0.991614	1.2557+/-0.09016	7.34E-07
rs16841224	2	209315198	T / C	0.083	0.9921	1.2560+/-0.09016	7.18E-07
rs34259292	2	209320832	T / G	0.1097	0.972886	1.2255+/-0.079968	6.42E-07
rs1490	5	87992873	G / A	0.7186	0.961407	0.8649+/-0.057624	8.15E-07
rs27643	5	87994702	T / G	0.7184	0.965215	0.8634+/-0.057624	5.68E-07
rs27721	5	87997279	A / G	0.7408	0.980215	0.8606+/-0.058408	4.82E-07
rs384005	5	88005103	T / C	0.7416	0.994048	0.8598+/-0.058212	3.56E-07
rs373098	5	88007261	C / T	0.7435	0.996408	0.8604+/-0.058212	4.13E-07
rs254777	5	88008871	G / C	0.7435	0.99671	0.8604+/-0.058212	4.16E-07
rs6475417	9	20212041	A / G	0.3949	0.977822	1.1434+/-0.05292	6.71E-07
rs321227	9	20215407	G / A	0.6103	0.979236	0.8747+/-0.05292	7.34E-07
rs321225	9	20220838	C / T	0.6043	0.973139	0.8751+/-0.05292	7.85E-07
rs321223	9	20222834	C / T	0.6037	0.971209	0.8754+/-0.05292	8.45E-07

[1] Odds Ratios were calculated with respect to A2.

Supplementary Table 3. GWAS estimates for SNPs linked to suicidal ideation or suicide attempt by previous studies.

Study	SNP	Chromosome	Chromosome position	Minor/major allele	Original study: level of significance	Current study: Odds ratio [CI-95%][1]	Current study: Level of significance
Perroudet al., 2010							
	rs358592	4	21475367	C / T	3.00E-06	0.9909+/-0.04390	6.85E-01
	rs4732812	8	28065871	C / T	3.00E-06	1.0071+/-0.04430	7.55E-01
	rs11143230	9	72272787	A / C	7.00E-06	1.0153+/-0.04214	4.80E-01
	rs11143230	9	72272787	A / C	8.00E-07	1.0153+/-0.04214	4.80E-01
Perlis al., 2010							
	rs2462021	10	30201775	C / T	8.00E-06	0.9692+/-0.04175	1.43E-01
	rs4918918	10	95362484	T / C	3.00E-06	1.0124+/-0.04136	5.59E-01
	rs10854398	21	39649825	T / C	6.00E-06	1.013+/-0.04018	5.29E-01
	rs12373805	2	46093955	G / A	9.00E-06	0.9631+/-0.04978	1.39E-01
Willour al., 2011							
	rs10437629	11	33566664	A / G	4.00E-06	0.9121+/-0.08350	3.07E-02
	rs7296262	12	128610527	T / C	1.00E-06	1.0104+/-0.04057	6.16E-01
	rs300774	2	112496	A / C	5.00E-08	0.9978+/-0.05351	9.35E-01
Mullins al., 2014							
	rs10748045	12	66422359	A / G	1.00E-06	0.9959+/-0.04155	8.45E-01
	rs3781878	11	113249490	A / G	2.00E-06		
	rs17387100	4	15993502	A / G	8.00E-07	0.9556+/-0.07742	2.50E-01
	rs17173608	7	150339575	T / G	2.00E-07	1.0352+/-0.09016	4.51E-01
Zai al., 2014							
	rs7079041	10	32704340	G / A	2.00E-06	0.9925+/-0.04253	7.31E-01
	rs7244261	18	68547459	C / T	4.00E-06	1.0105+/-0.05312	7.01E-01
	rs2610025	8	56592754	C / A	5.00E-06	1.0218+/-0.04057	2.97E-01
	rs10448044	8	79191197	T / C	3.00E-06	1.0099+/-0.04763	6.84E-01
Galfalvy al., 2015							
	rs320461	1	213424498	A / G	4.00E-06		
	rs336284	7	35254361	A / G	2.00E-07	0.9797+/-0.04077	3.25E-01

rs7011192	8	10677867	A / G	4.00E-06		
rs4308128	2	115733670	A / C	4.00E-06	1.0173+/-0.04038	4.04E-01
rs6480463	10	70758081	C / T	2.00E-06	0.993+/-0.04096	7.36E-01
rs4575	14	24146226	A / G	8.00E-06		
rs11852984	15	34813456	A/C/T	2.00E-06		
rs336284	7	35254361	A / G	8.00E-06	0.9797+/-0.04077	3.25E-01
rs3019286	8	98883177	G / A	8.00E-06	1.0114+/-0.04096	5.87E-01
rs2419374	1	190088550	C / T	1.00E-06	0.9973+/-0.05508	9.23E-01
rs6055685	20	8233139	G / A	8.00E-07	1.0155+/-0.05174	5.59E-01
rs13358904	5	112267016	A / G	5.00E-06	0.9598+/-0.04861	9.79E-02
Stein al., 2017						
rs2497117	6	84770179	A / G	1.58E-08	1.0954+/-0.10447	8.78E-02
rs2497118	6	84771964	A / G	1.70E-08	1.1018+/-0.10662	7.44E-02
rs2497119	6	84772961	A / C	1.18E-08	1.1061+/-0.10525	6.04E-02
rs2480192	6	84772469	T / C	1.32E-08	1.0925+/-0.10290	9.19E-02
rs142060512	6	84794805	C / T	3.55E-08	1.1098+/-0.16288	2.10E-01
rs116878613	6	84820786	T / C	4.12E-09	1.779+/-0.18659	8.56E-02
rs117975834	6	84898516	G / C	2.12E-08	1.1189+/-0.15327	1.51E-01
rs78022606	6	84914920	G / A	4.14E-08	1.0418+/-0.11309	4.78E-01
rs12524136	6	84935441	C / T	5.24E-10	1.053+/-0.12720	4.26E-01

Abbreviations: NR: Not reported.

[1] Model 2 was adjusted for: gender, number of years under follow-up being (aged ≥15 years), first 10 principal components of genetic ancestry, diagnosis of any mental disorder as well as diagnosis of schizophrenia, bipolar disorders, affective disorders, autism spectrum disorders, anorexia, and ADHD.

DECLARATION OF CO-AUTHORSHIP

The declaration is for PhD students and must be completed for each conjointly authored article. Please note that if a manuscript or published paper has ten or less co-authors, all co-authors must sign the declaration of co-authorship. If it has more than ten co-authors, declarations of co-authorship from the corresponding author(s), the senior author and the principal supervisor (if relevant) are a minimum requirement.

1. Declaration by	
Name of PhD student	Vivek Appadurai
E-mail	vivek.appadurai@regionh.dk
Name of principal supervisor	Dr. Thomas Werge
Title of the PhD thesis	Genetic Analysis of Complex Traits in Population Scale Datasets

2. The declaration applies to the following article	
Title of article	Genetics of suicide attempts in individuals with and without mental disorders: a population based genome wide association study
Article status	
Published ■ Date: 16/08/2018	Accepted for publication ☐ Date:
Manuscript submitted ☐ Date:	Manuscript not submitted ☐
If the article is published or accepted for publication, please state the name of journal, year, volume, page and DOI (if you have the information).	https://doi.org/10.1038/s41380-018-0218-y Molecular Psychiatry, 2018, 25, 2410-2421

3. The PhD student's contribution to the article ***(please use the scale A-F as benchmark)*** Benchmark scale of the PhD-student's contribution to the article **A.** Has essentially done all the work (> 90 %) **B.** Has done most of the work (60-90 %) **C.** Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
1. Formulation/identification of the scientific problem	C
2. Development of the key methods	B
3. Planning of the experiments and methodology design and development	B
4. Conducting the experimental work/clinical studies/data collection/obtaining access to data	B

3. The PhD student's contribution to the article ***(please use the scale A-F as benchmark)*** Benchmark scale of the PhD-student's contribution to the article **A.** Has essentially done all the work (> 90 %) **B.** Has done most of the work (60-90 %) **C.** Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
5. Conducting the analysis of data	A
6. Interpretation of the results	B
7. Writing of the first draft of the manuscript	C
8. Finalisation of the manuscript and submission	C

Provide a short description of the PhD student's specific contribution to the article.[i]

Contributed to study design, quality control of genotypes. Contributed to the methodological design of the analysis. Performed phasing and imputation of the data. Performed data analysis including estimates of heritability, genome wide association studies and annotation of results. Wrote the methods and results sections and contributed to the draft of the final manuscript.

4. Material from another thesis / dissertation[ii]	
Does the article contain work which has also formed part of another thesis, e.g. master's thesis, PhD thesis or doctoral dissertation (the PhD student's or another person's)?	Yes: ☐ No: ■
If yes, please state name of the author and title of thesis / dissertation.	
If the article is part of another author's academic degree, please describe the PhD student's and the author's contributions to the article so that the individual contributions are clearly distinguishable from one another.	

5. Signatures of the co-authors[iii]

	Date	Name	Title	Signature
1.	2.12.20	Annette Erlangsen	PhD	
2.	11.12.20	Andrew J Schork	PhD	

5. Signatures of the co-authors[iii]				
3.	14/12-2020	Thomas Werge	PhD	
4.	2/12-20	Merete Nordentoft	PhD	
5.		Esben Agerbo	PhD	
6.				
7.				
8.				
9.				
10.				

6. Signature of the principal supervisor
I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge. Date: 14/12-2020 Principal supervisor:

7. Signature of the PhD student
I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge. Date: 10-12-2020 PhD student: VIVEK APPADURAI

Please learn more about responsible conduct of research on the Faculty of Health and Medical Sciences' website.

[i] This can be supplemented with an additional letter if needed.

[ii] Please see Ministerial Order on the PhD Programme at the Universities and Certain Higher Artistic Educational Institutions (PhD Order) § 12 (4):

"Any articles included in the thesis may be written in cooperation with others, provided that each of the co-authors submits a written declaration stating the PhD student's or the author's contribution to the work."

[iii] If more signatures are needed please add an extra sheet.

PAPER 2

Aarøe L[a], **Appadurai V[a]**, Hansen K.M, Schork AJ, Werge T, Mors O, Børglum AD, Hougaard DM, Nordentoft M, Mortensen PB, Thompson WK, Buil A, Agerbo E, Petersen MB. **Genetic predictors of educational attainment and intelligence test performance predict voter turnout**. Nat Hum Behav (2020). https://doi.org/10.1038/s41562-020-00952-2

[a] Shared first authorship

ARTICLE

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TABLES

DECLARATION OF CO-AUTHORSHIP

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-00952-2

Genetic predictors of educational attainment and intelligence test performance predict voter turnout

Lene Aarøe[1,15], Vivek Appadurai[2,3,15], Kasper M. Hansen[4], Andrew J. Schork[2,3], Thomas Werge[3,5], Ole Mors[3,6], Anders D. Børglum[3,7], David M. Hougaard[3,8,9], Merete Nordentoft[3,10], Preben B. Mortensen[3,11,12], Wesley Kurt Thompson[3,13], Alfonso Buil[3], Esben Agerbo[3,14] ✉ and Michael Bang Petersen[1]

Although the genetic influence on voter turnout is substantial (typically 40–50%), the underlying mechanisms remain unclear. Across the social sciences, research suggests that 'resources for politics' (as indexed notably by educational attainment and intelligence test performance) constitute a central cluster of factors that predict electoral participation. Educational attainment and intelligence test performance are heritable. This suggests that the genotypes that enhance these phenotypes could positively predict turnout. To test this, we conduct a genome-wide complex trait analysis of individual-level turnout. We use two samples from the Danish iPSYCH case–cohort study, including a nationally representative sample as well as a sample of individuals who are particularly vulnerable to political alienation due to psychiatric conditions ($n = 13{,}884$ and $n = 33{,}062$, respectively). Using validated individual-level turnout data from the administrative records at the polling station, genetic correlations and Mendelian randomization, we show that there is a substantial genetic overlap between voter turnout and both educational attainment and intelligence test performance.

Representative democracies are premised on the electoral participation of their citizens. Without politically engaged citizens, democracy lacks its central impetus, and if some groups are systematically excluded from the political process or decide to opt out, key societal problems will most likely remain unaddressed, because the parliaments do not represent the views and priorities of their citizens. However, although the act of voting in this sense can be considered to be a fundamental public good, a large body of literature documents strong disparities among citizens in their level of electoral participation; some do not vote in elections at all[1–4]. These disparities exist both in the general population and in segments that are particularly vulnerable to political exclusion and alienation (such as those with psychiatric conditions). Why do some people devotedly vote in elections while others consistently remain unengaged?

A large body of twin-based studies has shown that individual differences in voter turnout are strongly associated with genetic variation, with heritability estimates that are typically between 40–50% but up to 72% (refs. [5–9]). Importantly, these findings raise the fundamental question of how genes and voter turnout are linked[5,10]. As emphasized by social scientists, "[a]t this early point in the research, there remains a black box"[11]. Together with a call that better theorizing 'must occur'[12], critiques have also been raised against existing twin and candidate gene studies owing to methodological limitations[13,14].

To advance our understanding of how genes and voter turnout are linked, we followed recent studies[5,10] and integrated the genetic and the social science paradigms of the microlevel foundations of voter turnout. Classic social science models of electoral participation emphasize that voting imposes opportunity costs on voters in the form of time, money, effort and cognitive investment[15–17]. In the face of these costs, individual differences in educational attainment and intelligence test performance constitute key predictors of political participation because they index 'resources for politics' that reduce the costs of voting. The correlations predicted by the resource model between political participation and educational attainment and cognitive performance, respectively, have found wide empirical support[17,18], and can also be observed in Denmark, the site of this study[19] (Supplementary Appendix 2.3 and Supplementary Table 5). However, from a genetics perspective, it is important to note that educational attainment and performance on intelligence tests are themselves genetically influenced and correlated traits[20–24]. Thus, the genetic variation that predicts individual differences in educational attainment and intelligence test performance may also

[1]Department of Political Science, Aarhus University, Aarhus, Denmark. [2]Institute of Biological Psychiatry, Mental Health Center Sct Hans, Roskilde, Denmark. [3]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus, Denmark. [4]Department of Political Science, University of Copenhagen, Copenhagen, Denmark. [5]Institute of Biological Psychiatry, Copenhagen Mental Health Services, Copenhagen, Denmark. [6]Psychosis Research Unit, Aarhus University Hospital—Psychiatry, Aarhus, Denmark. [7]Department of Biomedicine and Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark. [8]Center for Genomics and Personalized Medicine, Aarhus University, Aarhus, Denmark. [9]Danish Center for Neonatal Screening, Statens Serum Institut, Copenhagen, Denmark. [10]Mental Health Centre Copenhagen, Capital Region of Denmark, Copenhagen University Hospital, Copenhagen, Denmark. [11]NCRR-National Center for Register-Based Research, Business and Social Sciences, Aarhus University, Aarhus, Denmark. [12]CIRRAU–Centre for Integrated Register-Based Research, Aarhus University, Aarhus, Denmark. [13]Department of Family Medicine and Public Health, University of California, San Diego, CA, USA. [14]Department of Economics and Business Economics—National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. [15]These authors contributed equally: Lene Aarøe, Vivek Appadurai. ✉e-mail: ea@econ.au.dk

predict variation in voter turnout. By integrating the genetic and social science paradigms regarding the roots of voter turnout, we therefore investigated the following question: to what extent is the genetic variation that underlies educational attainment and intelligence test performance associated with individual differences in voter turnout? We argue and demonstrate that the genotypes that predict individual differences in educational attainment and intelligence test performance also predict individual differences in voter turnout.

For decades, scholars have especially considered education to be the 'universal solvent' that predicts high political participation and, therefore, the functioning of democracy (p. 324 in[25]; see also refs. [4,17,26,27]). Education "increases civic skills and political knowledge", which function as mechanisms underlying participation[28], and these mechanisms are associated with participation through "knowledge about where, when and for whom to vote"[6], the ability to communicate effectively, cognitive skills to understand policies, a sense of civic duty and responsibility[4,17], and the socioeconomic status that provides access to decision makers and politically relevant information[26,29]. Educational attainment is also associated with personality traits, such as openness to experience, that predict individual awareness and interest in politics[10,12,30], and previous research has argued that a higher educational achievement could also reflect an individual's motivation and self-control[31,32]. Importantly, these cognitive and non-cognitive mechanisms all reduce the individual's voting costs[17]. The causal interpretation of these relationships remains an ongoing discussion. In the genetics literature, research has found that the single-nucleotide polymorphism (SNP) associations that predict variation in educational attainment and cognitive performance also, to some extent, predict personality traits[33], socioeconomic status[34,35] and wealth[36]. In line with these results, there is ongoing debate in the political science literature regarding the causal interpretation of the relationship between phenotypic educational attainment and voter turnout and the underlying mechanisms (reviewed previously[28,37]). Persson[28] concluded that education might capture an individual's cognitive skills, motivation and social network. The question of the exact causal pathways notwithstanding, "[t]he relationship between education and voter turnout ranks among the most extensively documented correlations"[37], and educational attainment is considered to be one of the most important predictors of political participation[4,17,27,28].

In addition to educational attainment, cognitive performance indicators are also studied to predict voter turnout[38,39]. Although intelligence test performance should be interpreted with care, intelligence test performance has been viewed as an indicator of cognitive skills (discussed in ref. [40]) and a reliable predictor of political efficacy, which is associated with higher political participation[41]. Studies in western countries have found that those who perform well on intelligence tests during childhood are more interested in politics in adulthood and are more likely to participate in elections and engage in other forms of political participation[38]. Importantly, similar to education, it has also been found that intelligence test performance indicators are correlated with socioeconomic status[42,43], which is also associated with higher political participation[17].

Intelligence test performance and educational attainment are highly heritable and correlated polygenic traits ($r_g = 0.73$)[44,45]. The trait-enhancing markers are spread throughout the human genome[46–48] (reviewed previously[49]), with estimates of 51% narrow-sense heritability for intelligence[50] and 21% for educational attainment[51] obtained from the genomic-relatedness-based restricted maximum-likelihood approach implemented in genome-wide complex trait analysis (GCTA-GREML). Thus, integrating the social science resource model of political participation with the genetic literature on the heritability of educational attainment and intelligence test performance gives rise to the key correlational hypothesis that the genetic variance that predicts educational attainment and intelligence test performance, respectively, predicts individual differences in electoral participation.

In testing this hypothesis, we also sought to overcome methodological shortcomings in the existing literature. Previous studies investigating the heritability of voter turnout have relatively low external validity owing to an almost exclusive reliance on samples of twins ($n = 396$–$3{,}616$)[5–10]. One molecular genetics study relied on a candidate-gene approach conducted on a random sample of participants in the United States[52], but concerns have been raised about its replicability[14]. A recent study performed the first genome-wide association study (GWAS) for voting behaviour, but used regional data on voting behaviour and did not analyse individual-level turnout[53]. All but three studies have used data from individuals from the United States, which further limits the external validity. Interestingly, one of these studies with Swedish participants produced mixed findings regarding the heritability of voting[5]. Furthermore, it is important to note that all research efforts have been directed towards investigating the heritability of voting[5–9,52] and the underlying mechanisms[5] in the general population. Investigation into the genetics of individual-level turnout in vulnerable subpopulations in which political non-participation is a more wide-spread phenomenon than in the general population is lacking[54,55] (Supplementary Appendix, Supplementary Table 1 and Supplementary Fig. 1).

Previous studies addressing the heritability of voter turnout also face issues regarding internal validity. Most previous studies used the classic reared-together twin design, which typically confounds the potential influence of genes and shared environment (discussed previously[56]). Given these limitations to the internal validity, scholars have warned that twin studies "can neither prove nor refute the argument for a genetic basis of political traits such as […] voting turnout"[57]. Finally, most studies rely on self-reported voter turnout, which is known to be associated with misreporting and over-reporting and, therefore, limits measurement validity[58].

To address these limitations to external, internal and measurement validity, we conducted a comprehensive investigation of the link between genetics and electoral participation. To increase internal and external validity, we moved beyond the classic twin design and relied on a dataset of comprehensively genotyped individuals from the Danish-population-based iPSYCH case–cohort sample[59]. It includes a representative sample of the entire population born between 1981–2005 in Denmark and a psychiatric sample of people diagnosed with anorexia, schizophrenia, affective disorder (including depression), bipolar affective disorder, autism and attention-deficit/hyperactivity disorder (ADHD) born in Denmark during the same period[59] (Supplementary Appendix 1).

As the overall setting, the Danish electoral context is substantially more inclusive than the United States, where most previous studies of the heritability of turnout have been conducted and where, for example, comprehensive voter identification laws shape turnout[60]. In Denmark, all eligible citizens are automatically registered to vote, and much is done to enable disabled people and others on the margins of society to vote in elections. The barriers to electoral participation are therefore relatively low in Denmark.

Individuals with a psychiatric condition are an important subpopulation to study. According to the WHO, one in four people in the world become affected by a psychiatric or neurological condition during their lives[61]. Individuals with a psychiatric condition are potentially vulnerable to social and political exclusion[54,62–67]. First, we might expect that psychiatric condition, on average, will lower turnout. Using the logic of the resource model, previous research therefore argues that poor health may "hinder the acquisition of other resources such as civic skills, time and money"[66]. Second, we might expect that individuals with certain psychiatric conditions are particularly vulnerable to exclusion. For example, previous studies have argued that impeded cognitive functioning—which is central in psychiatric diagnoses such as ADHD—"negatively

influence[s] individuals' ability to process information related to elections"[55]. Having a psychiatric condition may also lead people to actively deprioritize voting as too costly in terms of time, effort and attention and therefore voluntarily opt out of the political process. Finally, other studies argue that some types of mental conditions such as depression can reduce "the somatic capacity of an individual and therefore reduce the resources an individual has for political participation"[54]. If individuals with psychiatric conditions do not vote in elections, parliaments may not represent the views and priorities of this group of citizens, and this augments the risk that key problems and needs relevant to them will remain underprioritized.

In our psychiatric sample, individuals with a psychiatric diagnosis were on average 15–32% less likely to vote in the three elections under study than the general Danish population (Supplementary Appendix 2, Supplementary Table 1 and Supplementary Fig. 1). Individuals who were diagnosed with anorexia have the highest turnout frequency and are descriptively indistinguishable from the representative sample. Individuals who were diagnosed with ADHD have the lowest turnout and are 28–54% less likely to vote than the average Danish population (Supplementary Appendix 2, Supplementary Table 1 and Supplementary Fig. 1). Studying the link between genes and voter turnout in both a nationally representative sample and a subpopulation of individuals with a psychiatric diagnosis who are vulnerable to political exclusion increases the external validity of the results.

To maximize measurement validity, we moved beyond self-reports of turnout and integrated the genetic data with population-based turnout data obtained directly from the administrative records at the polling place of whether Danish citizens voted in the 2015 national, 2014 European and 2013 municipal elections (Supplementary Appendix 1.2). The saliency of the election differs substantially across these types of elections, most simply apparent from the general turnout in the elections. The national election constitutes a high-saliency election with a turnout of 86% in 2015. The municipal election in 2013 represents a medium-high-saliency election with a turnout of 72%. Finally, European elections generally have low saliency in Denmark, and turnout in 2014 was only 56%. Investigating our prediction across three types of elections further increases external validity.

With 46,946 unrelated individuals, this provides a dataset that excels in terms of (1) sample size; (2) measurement validity, as electoral participation is objective rather than self-reported; (3) internal validity, as the research design moves beyond the limitations of the classic reared-together twin design; and (4) external validity, as the dataset includes a random, nationally representative sample of more than 13,000 individuals and one of the world's largest samples of a subpopulation of individuals with a psychiatric diagnosis, and covers three types of elections.

Results

Using GCTA, we first estimated the proportion of phenotypic variation in electoral participation in each of the three elections that can be accounted for by SNPs (hereafter refered to as h^2_{SNP}), which are the most common type of genetic variations between individuals. Previous estimates from twin and family studies typically conflate heritability because they are "biased by factors shared by close relatives, such as non-additive genetic and common environmental effects"[68,69]. The h^2_{SNP} is not inflated by these confounds (h^2_{SNP} captures only variance explained by common SNP variants). It can therefore be seen as an estimate of the lower bound of the total (or broad sense) heritability (H^2). The h^2_{SNP} estimates are reported in Fig. 1 with 95% confidence intervals (CIs), enabling us to compare the effect of heritability between samples (national and psychiatric) and across the types of election (varied by the election saliency).

In the psychiatric sample, we found a statistically significant h^2_{SNP} for electoral participation phenotypes with the point estimates, suggesting that the common SNPs explain approximately 8–10% (s.e. = 0.0155–0.021) of the phenotypic variation in electoral participation (all $P \leq 2.61 \times 10^{-7}$). Previous American twin studies of the heritability of electoral turnout have estimated that up to 53–72% of individual differences in electoral turnout can be accounted for by genetic variability. Previous studies focusing on height[70], intelligence[50], personality traits[71], and political and economic preferences[72] found that h^2_{SNP} estimates are approximately 1/2 to 1/4 the size of twin-study estimates. On this basis, the observed SNP-based heritability estimates in the psychiatric sample are proportionally consistent with previous American twin-study estimates.

In the nationally representative sample, the h^2_{SNP} estimate for the 2014 European election appears to be qualitatively similar to the estimates from the psychiatric sample ($h^2_{SNP} = 0.1096$, s.e. = 0.0044, $P = 6.43 \times 10^{-3}$; Fig. 1a). The h^2_{SNP} estimates for voting in the in the high- to medium-high-saliency elections in the 2015 and 2013 elections are statistically non-significant, and the point estimates are lower compared to the psychiatric sample. Importantly, the substantial overlap between the CIs for the heritability estimates in the two samples indicates that the h^2_{SNP} estimates in the nationally representative sample are not statistically significantly different compared to the psychiatric sample, and that the overall trends are the same in the two samples. Part of the explanation for the non-significant h^2_{SNP} estimates in the nationally representative sample is the lower statistical power in this sample (a power analysis is provided in Supplementary Appendix 3 and Supplementary Fig. 2) and that the h^2_{SNP} estimate represents a lower bound for narrow sense heritability. To further explore this interpretation, we repeated the analysis in Fig. 1 using pedigree-defined familial relationships. These analyses provide highly statistically significant ($P < 1.24 \times 10^{-40}$) upper-bound heritability estimates of voter turno\ut that range between 0.39 (s.e. = 0.02) and 0.49 (s.e. = 0.02) across the three elections (Supplementary Appendix 4.2 and Supplementary Tables 6.1 and 6.2). Although the family estimates may be biased upwards due to the expected common environment effect in a trait such as voting, the results overall support that the statistically non-significant h^2_{SNP} estimates in the nationally representative sample may be due to the lower statistical power in this sample and that the h^2_{SNP} estimate represents a lower bound for the narrow sense heritability.

Interestingly, in the nationally representative sample, the SNP heritability is significantly higher in the low-saliency European parliamentary election compared with in the national election ($P = 0.044$, $z = 1.7052$). Although the difference in point estimates is similarly wide in the municipal elections, it is not statistically significant owing to a larger s.e. of the estimate ($P = 0.0534$, $z = 1.6126$); details of these analyses are provided in the 'Analyses of differences in heritability across elections' section in the Methods). The statistically significant difference in the SNP heritability in the low-saliency European parliamentary election compared with the high-saliency national election is consistent with Tingsten's[73] law of dispersion, which predicts that the participatory gap between high- and low-propensity voters is higher in low-saliency elections because low saliency increases the costs of voting[74]. In such circumstances, the effect of individual differences including genetic dispositions should take precedent, which is the pattern that we observed. This perspective could also potentially explain the general low heritability estimates in the Danish nationally representative sample compared with previous studies, given the general inclusiveness of the Danish electoral context. Regarding the psychiatric sample, note that we found no statistically significant difference between SNP heritability of turnout in the European election of 2014 and in the national election of 2015 ($P = 0.4099$, $z = 0.2277$) or the municipal election of 2013 ($P = 0.2561$, $z = 0.6551$); details of these analyses are provided in the 'Analyses of differences in heritability across elections' section in the Methods). Thus, the heritability of turnout is possibly less affected by election saliency in this population.

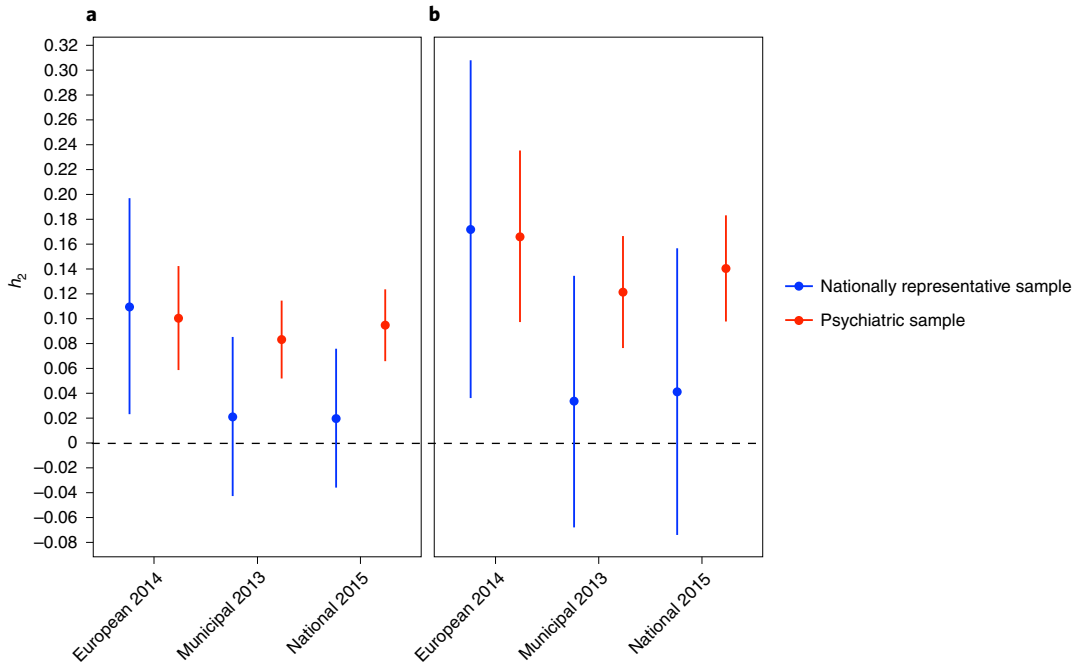


Fig. 1 | Heritability estimates for voting in municipal, European and national elections. **a**,**b**, Observed scale (**a**) and a transformed liability scale (**b**). Data points refer to point estimates for narrow-sense SNP heritability. Error bars incidcate 95% CIs.

Predictive use of polygenic scores. Next, we investigated whether the genotypes that predict individual differences in educational attainment and performance on intelligence tests also predict individual differences in voter turnout. We used polygenic scores (PSs) to measure genetic correlates of educational attainment and performance on intelligence tests (compare with refs. [44,75,76]). A PS is an index that aggregates the effects of all of the DNA variants associated with a given trait to predict an outcome[24]. Specifically, a PS is the sum of the products of additive counts of SNP alleles (for example, 0, 1 or 2 copies of the rarer of two alleles at a given variant) and the corresponding per-allele effect sizes estimated on the basis of a GWAS for a given trait. PSs have been shown to reliably predict educational attainment[75] and intelligence test performance[44]. However, the predictive ability of PSs should be interpreted knowing that the proportion of explained variation is typically low[77]. About 11–13% of the variance in education and 7–10% of the variance in cognitive performance can be predicted from the PS for educational attainment, and 5.2% of the variation in intelligence test performance can be predicted from the PS for intelligence test performance[44,75].

Figure 2 shows voter turnout within each percentile of the distribution for the PS for educational attainment and performance on intelligence tests across the three elections in the psychiatric and nationally representative samples. The results show that higher PSs for educational attainment and intelligence test performance correspond with increased voter turnout across all three elections in both samples. Details of the variance in voter turnout explained by PSs for educational attainment and intelligence test performance for each election across both the nationally representative and psychiatric samples and the associated P values are provided in Supplementary Appendix 5.2 and Supplementary Table 7. These results indicate that the relationships between voter turnout and PS for educational attainment and intelligence test performance, respectively, are highly statistically significant ($P < 1.61 \times 10^{-7}$ or lower) and that the gain in Nagelkerke pseudo r^2 ranges from 0.0041 (95% CI = 0.0016–0.0078) to 0.0317 (95% CI = 0.0235–0.0413) when one of the PSs was added to a baseline logistic model that had the first ten principal components of genetic ancestry as explanatory variables (Supplementary Appendix 5.2 and Supplementary Table 7). Although voter turnout is overall higher in the nationally representative sample, the predictive value of the PSs retains statistical significance (Fig. 2 and Supplementary Table 7). The consistency in predictive performance of the PSs in the two samples suggests that the overlap among SNPs associated with voter turnout and educational attainment or intelligence test performance is comparable in the general population and in the subpopulation of individuals with a psychiatric condition.

When interpreting the phenotypic processes producing these findings, it is relevant to consider previous studies that showed that SNPs associated with educational attainment may in part exert their influence through personality traits[78]. Social science studies have also found evidence of a relationship between the big five personality traits and political participation[12,30,79–81] (reviewed previously[82]). In particular, some studies suggest that personality factors can account for part of the correlation between genetics and political participation[5,10] and highlight extraversion as a factor that "may possibly mediate the relationship between genes and political predispositions that are known to be strongly related to political participation"[5].

Although we believe that it is probable that phenotypic personality constitutes part of the causal pathway from genetics to turnout, we cannot test whether the correlation between PSs for educational attainment and turnout is mediated by the phenotypic big five personality traits or phenotypic education, as these variables are not available in our data. However, further analyses reported in Supplementary Appendix 5.2 and Supplementary Tables 7 and 8 show that the PSs for educational attainment or intelligence test performance predict individual differences in turnout more consistently than the PSs for the big five personality traits. The PSs for the big five personality traits display mixed predictive utility for electoral turnout in which the gain in Nagelkerke pseudo r^2 ranges between 2×10^{-10} and 0.0014 when the PSs were added to a baseline logistic regression model that had the first ten principal components of genetic ancestry as explanatory variables, and P values for the statistical significance of the PSs for the big five personality traits range

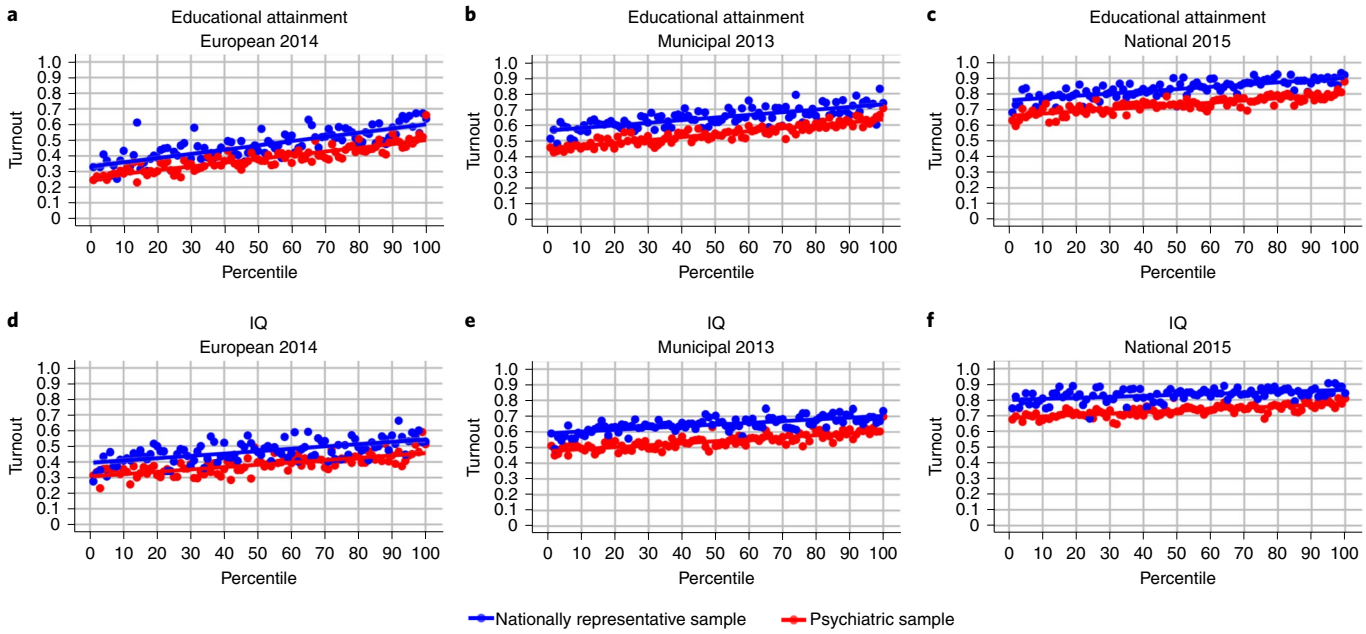


Fig. 2 | Prevalence of voter turnout by polygenic score percentiles. a–f, Voter turnout within each percentile of the distribution for the PSs for educational attainment (**a**–**c**) and intelligence test performance (**d**–**f**) for the European (**a**,**d**), municipal (**b**,**e**) and national elections (**c**,**f**).

between 0.9754 and 4.32×10^{-5} (Supplementary Appendix 5.2 and Supplementary Table 8). For example, across the three elections and our two samples, only in the 2014 European election in the psychiatric sample did we find that the PS for extraversion significantly accounts for variation in turnout. PS for openness to experience also has mixed predictive use and no significant predictive utility of conscientiousness were observed. Only the PS for agreeableness is consistently significantly associated with turnout across the three elections in the two samples. Note that these results could reflect the lower sample size of the dataset that was used to generate the genome-wide summary statistics for calculating the PS for the big five personality traits. Still, these findings support the more consistent predictive use of PS for educational attainment and intelligence test performance, respectively, compared to the PSs for the big five personality traits.

Genetic correlation analysis. Next, we investigated the genetic correlation between voter turnout and traits that are related to educational attainment and intelligence test performance. Furthermore, we analysed genetic correlations between voter turnout and socioeconomic factors as well as health indicators using linkage disequilibrium score regression (LDSC)[83] as implemented in LDHub, which makes it possible to assess genetic correlation with many traits using online tools[45]. Specifically, we used LDHub to test the correlation between 72 traits and voter turnout. Separating the correlations by election and sample, Fig. 3 shows the genetic correlations between voter turnout and non-voting traits for all traits with at least one significant correlation across elections, and sample after false-discovery rate (FDR) adjustment (the full list of genetic correlations and all estimates is provided in Supplementary Appendix 11).

In the psychiatric sample, we found positive genetic correlations for electoral turnout across all three elections with indicators of educational attainment (that is, college completion and years of schooling) and intelligence test performance (that is, childhood IQ). As expected, the correlations in the nationally representative sample follow the same pattern, albeit with higher P values and larger error bars, most likely due to the smaller sample size (an analyses of robustness is provided in Supplementary Appendix 7.1). None of the genetic correlations in Fig. 3 show any discernible difference between the two samples for any of the elections (Supplementary Appendix 7.2 and Supplementary Fig. 7).

Interestingly, negative socioeconomic factors—such as number of children, younger age at first childbirth and younger parental age at death—also show genetic correlations with voter turnout (Fig. 3). This suggests that, in addition to the PSs for education and intelligence test performance, a broad set of socioeconomic resources could connect genetics to electoral participation. Furthermore, Fig. 3 shows a negative correlation between voter turnout and adverse physical and mental health outcomes (for example, depressive symptoms, neuroticism and ADHD). The genetic correlations with depressive symptoms and ADHD symptoms underline the importance of these psychiatric conditions as predisposing factors for non-voting. Even among individuals for whom the additive sum of conferred risk does not cross the threshold to manifest as a psychiatric diagnosis, instances of significant correlation with voter turnout are identified for each of these traits.

Summary-statistics-based Mendelian randomization analysis. As the final test of our prediction, we performed a Mendelian-randomization-based analysis[84] of whether the genetic variance that predicts educational attainment and intelligence test performance, respectively, predicts voter turnout. Specifically, we used markers surpassing the threshold for genome-wide significance ($P \leq 5 \times 10^{-8}$) in association studies of educational attainment and intelligence test performance as instrumental variables. These genetic markers for educational attainment and intelligence test performance, respectively, are treated as exposures and voter turnout in each election as an outcome.

The results (Supplementary Appendix 8.2, Supplementary Table 12 and Supplementary Fig. 8) show that an individual in the nationally representative sample who has a genetic disposition to obtain education 1 s.d. higher than the population mean was predicted to be 2.66 (s.e. = 0.1064, $P = 3.76 \times 10^{-20}$) times more likely to vote in the municipal election in 2013, 3.14 (s.e. = 0.129, $P = 7.51 \times 10^{-19}$) times more likely to vote in the European Parliament election in

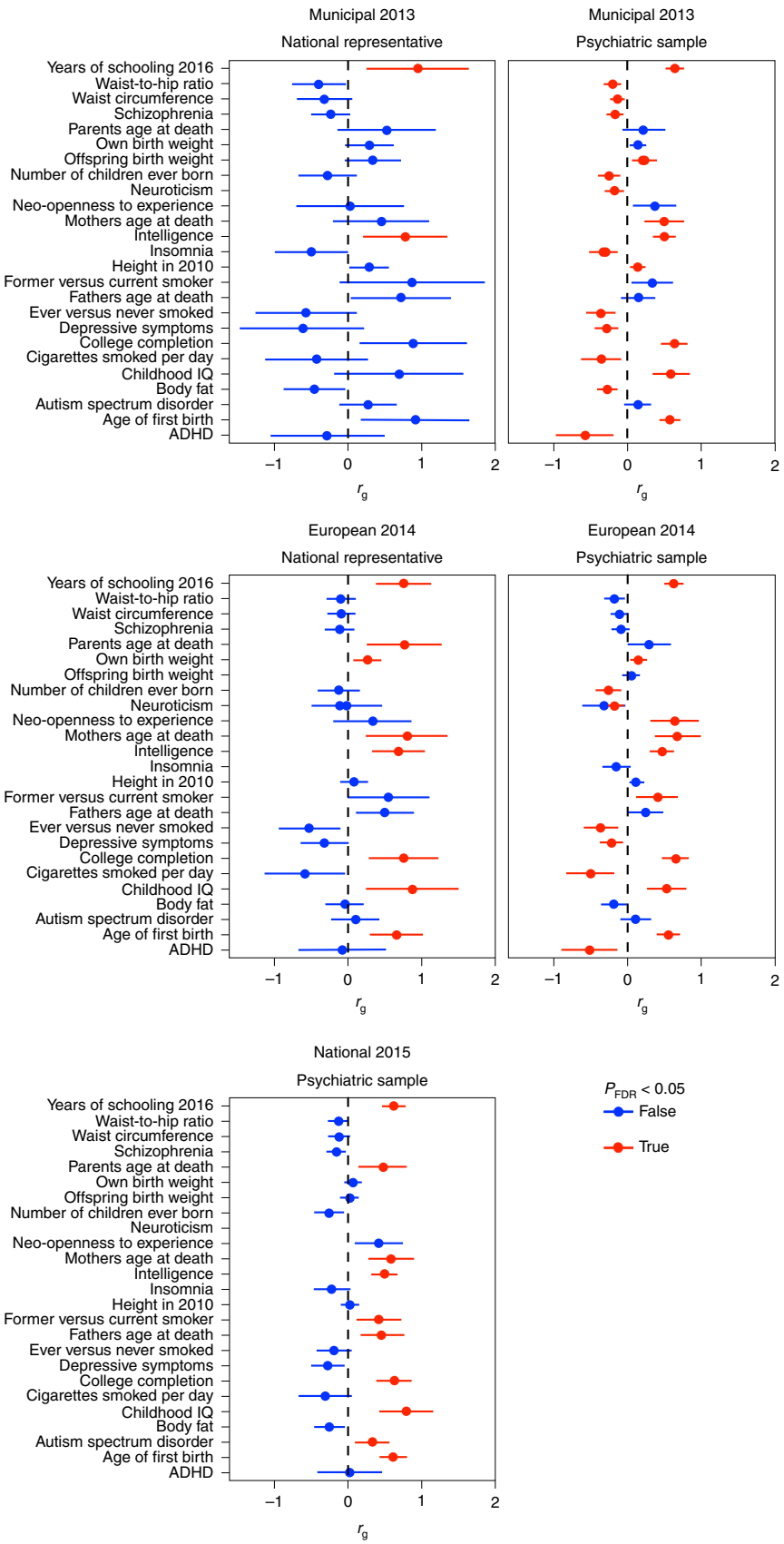


Fig. 3 | Genetic correlations between voter turnout and non-voting traits that show significant genetic correlation after FDR adjustment in at least one of the turnout phenotypes in the two samples by election. The LDSC-estimated heritability of turnout in the 2015 national election was not significantly different from zero in the nationally representative sample; therefore, no correlations were computed. Data points refer to point estimates for genetic correlations. Error bars incidcate 95% CIs.

2014 and 3.32 (s.e. = 0.1439, $P = 6.66 \times 10^{-17}$) times more likely to vote in the national election in 2015. An individual in the psychiatric sample with a genetic disposition to obtain education 1 s.d. higher than the mean in the psychiatric sample was predicted to be 2.97 (s.e. = 0.0866, $P = 3.44 \times 10^{-36}$) times more likely to vote in the 2014 European election, 2.65 (s.e. = 0.08, $P = 3.72 \times 10^{-34}$) times more likely to vote in the 2015 national election and 2.67 (s.e. = 0.0656, $P = 1.26 \times 10^{-50}$) times more likely to vote in the 2013 municipal election.

We observed a similar trend with intelligence test performance, although the magnitude of the effect size was smaller. An individual in the nationally representative sample with a genetic disposition to perform on intelligence tests 1 s.d. higher than the mean in the general population was predicted to be 1.85 (s.e. = 0.142, $P = 1.47 \times 10^{-5}$) times more likely to vote in the national election in 2015, 2.15 (s.e. = 0.1277, $P = 1.87 \times 10^{-9}$) times more likely to vote in the European Parliament election in 2014 and 1.77 times (s.e. = 0.1051, $P = 4.6 \times 10^{-8}$) more likely to vote in the municipal election in 2013. Similarly, an individual in the psychiatric sample with a genetic disposition to perform on intelligence tests 1 s.d. higher than the mean of this sample was predicted to be 1.78 (s.e. = 0.078902, $P = 2.58 \times 10^{-13}$) times more likely to vote in the national election in 2015, 2.07 (s.e. = 0.0858, $P = 1.19 \times 10^{-17}$) times more likely to vote in the European Parliament election in 2014 and 1.66 (s.e. = 0.0647, $P = 3.06 \times 10^{-15}$) times more likely to vote in the municipal election in 2013.

As a negative control, we repeated the analysis above using markers that surpass the genome-wide significance threshold for height obtained using summary statistics from the GIANT consortium as instrumental variables. The GSMR analysis of height did not suggest that there was any significant relationship between genetic dispositions towards height and turnout in any election across either of the samples (Supplementary Appendix 8.2 and Supplementary Table 12).

However, as our data are observational and Mendelian randomization analysis requires strong identifying assumptions and does not account for any environmental influences on the exposures of interest towards the outcome, the results should be seen as correlation evidence that PSs for educational attainment and intelligence test performance predict voter turnout (further discussion of our analysis and the assumptions for Mendelian randomization analyses are provided in Supplementary Appendix 9).

Conclusion

It is essential to the functioning and legitimacy of democracy that citizens vote in elections. Nonetheless, large inequalities in voter turnout exist. Although a decade of twin-based studies has shown that variability in voter turnout is shaped by genetics[5,7–9,13], the fundamental question of the mechanisms that underlie this connection largely remains a 'black box'[12]. Traditionally, social science models of voter turnout have "typically ignored genetic or biological factors"[85].

Here we followed recent studies[5,10] and sought to advance our understanding of how genes and voter turnout are linked by integrating the genetic and the social science paradigms of the microlevel foundations of voter turnout, arguing that the genetic variances that predict educational attainment and intelligence tests performance predict individual differences in voter turnout. To test this argument, we conducted a genome-wide complex trait analysis of the heritability of individual-level electoral participation using a nationally representative sample and a population sample of individuals diagnosed with anorexia, schizophrenia, affective disorder, bipolar affective disorder, autism and ADHD. Our methodological triangulation combining PSs, genetic correlations and Mendelian randomization analyses consistently supports that the genetic variants that predict educational attainment and intelligence test performance, respectively, also predict voter turnout.

While our findings provide evidence for the predicted genetic associations, we are unable to draw a causal interpretation. First, as we work with observational data and Mendelian randomization analysis requires strong identifying assumptions, our results are correlational. We cannot determine whether the observed genetic correlations reflect causal effects of phenotypic education and phenotypic intelligence test performance on voter turnout. For example, some research suggests that education works at least in part as a proxy for pre-adult factors, such as family socioeconomic status[86] (reviewed previously[28]), and recent research has found evidence of genetic nurturing in which genes influence educational attainment indirectly through the family environment[86].

Second, the resources for politics associated with, for example, educational achievement are multifaceted and include particular personality traits as well as material resources. We therefore find that it is probable that part of the relationship between genetic dispositions toward educational achievement or intelligence test performance could be mediated by phenotypic personality-related resources. Furthermore, as the PS for educational attainment is also predictive of performance on cognitive tests[75], part of the significant relationship between the PSs for educational attainment and turnout could be explained by the broader set of factors underlying such performance. Although our results are consistent with the argument that educational attainment and intelligence test performance represent a link between genes and voter turnout, respectively, they cannot provide conclusive evidence in support of such a causal interpretation.

The above limitations notwithstanding, our study contributes by moving beyond the limitations of the reared-together twin design to analyse the heritability of individual-level voter turnout using data on SNPs. Given the strong critiques raised in the social sciences of existing twin-based studies of the heritability of turnout[13,14], this is important as methodological triangulation helps to overcome some of the limitations of twin-based studies and increases internal validity. Considering the typical differences in the size of h^2_{SNP} estimates and twin-study estimates[87], the h^2_{SNP} estimates in the psychiatric sample are proportionally consistent with previous (American) twin-study estimates. Interestingly, we did not observe statistically significant h^2_{SNP} estimates in the nationally representative sample in the high- and medium-high-saliency elections in 2013 and 2015, respectively, but only in the low-saliency election for the European election in 2014. However, (1) the overlapping CIs with the psychiatric sample, (2) the power analysis and (3) the heritability estimates from the pedigree-defined familial relationships suggest that the non-significant h^2_{SNP} estimates for the high- and medium-high-saliency elections in the nationally representative sample in part reflect lower statistical power in this sample.

Importantly, the differences in heritability between elections and samples probably also reflect substantive contextual differences. First, in the nationally representative sample, we found higher heritability of turnout in the low-saliency European election compared with the high-saliency national election. Second, our results from the nationally representative sample suggest that the heritability for national and highly salient elections possibly could be lower in the general Danish population sample compared with previous estimates from United States; these findings are consistent with Tingsten's[73] law of dispersion, which stipulates that individual differences in resources for politics have stronger influence on whether people vote in elections with greater participatory costs as reflected in, for example, lower general turnout. Note that these results are based on data from a single country, that is, Denmark. Our results support the value of additional investigations of the differences in the heritability of turnout between elections with different costs of participation in different types of electoral systems to further explore the replicability and external validity of this finding.

Another path for further research is to study whether the heritability of turnout varies across psychiatric conditions. Our analysis focuses on average patterns for individuals with a psychiatric condition. However, certain types of psychiatric conditions may generate particularly high costs of voting. Drawing on the logic of Tingsten's law of dispersion[73], this could imply that these conditions increase the role of genetic dispositions.

Overall, the results extend understanding of the patterns that underlie inequalities in voter turnout. They show that the genetic variance that predicts educational attainment and intelligence test performance, respectively, also predicts individual differences in voter turnout. This integration of the social science and genetic paradigms of voter turnout reveals new linkages between social and political inequalities in the general population and in the subpopulations that are particularly vulnerable to political exclusion due to psychiatric diagnoses.

Methods

Data. The iPSYCH Danish case–cohort sample consists of two large samples of comprehensively genotyped individuals. A nationally representative sample ($n = 30{,}000$) of individuals born in Denmark during 1981–2005; and a population sample of all individuals born during the same period with one of the following psychiatric diagnoses: schizophrenia, affective disorder, bipolar affective disorder, autism, ADHD and anorexia ($n = 63{,}080$)[59].

The iPSYCH samples were merged with population-based records of whether Danish citizens voted in the 2015 national, 2014 European and 2013 municipal elections in Denmark. To maximize measurement validity, turnout data were collected directly from the voting files at the polling stations (Supplementary Appendix 1). We excluded individuals who show second degree or higher relatedness, individuals who were not from the genetically homogenous population, as identified by PCA analysis, and individuals who failed standard GWAS quality control tests for association tests. Finally, we limited the analysis to individuals who were aged 18 years or older and had therefore reached the Danish voting age to be eligible to vote in a public election in 2015 (Supplementary Appendix 1).

The iPSYCH has been approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, Statistics Denmark and the Danish Neonatal Screening Biobank Steering Committee[59]. The iPSYCH is "an example of Danish studies based on the combination of registers and biological material stored in biobanks"[88]. On 28 August 2012, in accordance with the Act on Research Ethics Review of Health Research Projects (in Danish, Komitéloven), Section 10 (1), the Central Denmark Region Committees on Health Research Ethics, Committee II granted the iPSYCH exemption from obtaining informed consent from participants (https://ipsych.dk/en/data-security/health-research-and-ethical-approval/). The committee's "processing and approval of the iPSYCH project was in accordance with applicable law and practice at the time of notification." (https://ipsych.dk/en/data-security/health-research-and-ethical-approval/). This waiver also applies to our study. Participants for the iPSYCH project were recruited according to consent by non-opt-out from a national research program[89,90]. Participants "can opt out of having any biological material used for research without specific informed consent"[88]. It is possible for participants to opt out of the Biobank storage at any time (further description of the briefing and opt-out procedures[90] and discussion of the scientific ethical standards in iPSYCH[88] were reported previously). The iPSYCH has data protection measures in place that comply with Danish and European legislation[88].

Measures. *Voting.* Voting in each election was coded as 1 = did vote and 0 = did not vote. All analyses include individuals who were eligible to vote at the time of the election and were Danish citizens.

PSs for educational attainment and intelligence test performance. PSs for educational attainment and intelligence test performance were calculated using summary statistics from large recent studies of associations between genetic markers and educational attainment in 766,345 individuals (excluding the 23andme cohort)[75] and intelligence in 269,867 individuals of primarily European ethnicity[44] (details on the base datasets for the PSs and the target dataset are provided in Supplementary Appendix 5.1).

Previous research indicates that the predictive ability of educational attainment varies across samples[75]. Only limited available data exist regarding the accuracy of PSs for educational attainment in Danish samples, but previous research suggests that PSs for educational attainment predicted about 15% of the variance in educational attainment in a small sample of 1,459 Danish individuals[75]. In terms of predictive accuracy, this is comparable to the size of the general estimate of about 11–13%.

PSs for the big five personality traits. PSs for the big five personality traits were calculated using summary statistics from large recent studies of associations between genetic markers and the big five personality traits[91] using summary statistics from 175,375 adult individuals, obtained from the Genetics of Personality Consortium (http://www.tweelingenregister.org/GPC/) as described previously[91].

To construct the PSs, we used PRSice v.1.25 (https://choishingwan.github.io/PRSice/) with the default parameters (clump-r2 = 0.1, clump-kb = 250, clump-p = 1) to sum the products of the effect size of each individual marker with the associated additive genotype for every individual in the iPSYCH determined to be of European origin using principal component analysis. Summary statistics for all measures and demographic sample characteristics are provided in Supplementary Appendix 2 and Supplementary Tables 1–4.

Statistical analysis. All of the reported tests of statistical significance were performed in a two-tailed manner.

Following standard practices, all analyses focus on genetically homogenous (that is, broadly northern European) ancestry respondents (compare with ref. [76]). A detailed description of quality control, haplotype estimation and genotype imputation protocols undertaken on the iPSYCH dataset is provided in Supplementary Appendix 1.1.2. In brief, the iPSYCH cohort of 78,050 individuals were genotyped at 554,360 genomic loci in 23 waves. After initial quality control was performed on the basis of principal component analysis using common (minor allele frequency > 5%) high-quality markers common to the iPSYCH and the full 1000 Genomes phase 3 variant calls, followed by principal component analysis and outlier detection, 75,501 samples of a homogenous genetic origin were selected for SNP-level quality control. After excluding rare markers (minor allele frequency < 1%), those that failed tests for Hardy–Weinberg equilibrium within controls and markers with high missingness, a total of 246,539 SNPs were phased using SHAPEIT3 and imputed to 80,707,375 SNPs in 10 batches using IMPUTE2 and the full set of 1000 Genomes phase 3 haplotypes as the reference. Following stringent post-imputation quality control to exclude SNPs with low imputation INFO scores, rare SNPs (minor allele frequency < 0.001), SNPs with high missingness, differential missingness between cases and controls and SNPs showing significant association to a genotyping wave or imputation batch, 11,601,089 SNPs were retained. Sample-level quality control involved excluding samples with greater than second-degree relatedness, samples showing discordance from documented gender, duplicate samples, samples with abnormal heterozygosity not explained by admixture, samples with high missingness and samples that deviated from a homogenous genetic background as identified by ancestral birth records in conjunction with principal component analysis. This resulted in a total of 65,535 samples that were used for all complex trait analyses. All of the reported statistical tests were performed in a two-sided manner, except where explicitly indicated.

GCTA. We controlled for sex, age (measured in days after the election and centred to the mean) and age^2 (centred to the mean), as this is the functional form in relation to turnout in a young cohort[92] and for the first ten principal components of the SNP-based genetic relatedness matrix. The genetic relatedness matrix (GRM) for estimating the GCTA-GREML SNP heritability was computed using the standard SNP filters chosen for iPSYCH data release (MAF > 0.001, imputation INFO score > 0.2)[89] and we used a grm-cutoff of 0.05 as a threshold for relatedness between samples. We conducted an analysis of robustness in which we built a GRM from SNPs common between the HapMap3 dataset[93] and iPSYCH with a minor allele frequency of ≥0.01 and imputation INFO scores of ≥0.8 and a more stringent relatedness threshold with a grm-cutoff of 0.034. Although the point estimates of SNP heritability vary slightly, the significance of our results do not change (Supplementary Appendix 4.1)

We used the publicly available GCTA software for the analysis[70]. When reporting the h^2 estimates for voting in the 2013, 2014 and 2015 elections in Fig. 1, we reported h^2_{SNP} estimates on both the observed scale from our phenotypes and on a transformed liability scale, in which the estimates are scaled according to the percentage of voters in the nationally representative sample.

Analyses of differences in heritability across elections. To examine whether the SNP heritability of turnout in the European election of 2014 is significantly higher than in the national election of 2015 or in the municipal election of 2013 in both the nationally representative and psychiatric samples, we computed the z score of the difference $z = h^2_{\text{European election 2014}} - h^2_{\text{municipal election 2013}}$ OR $h^2_{\text{national election in 2015}} / \sqrt{(\text{s.e.}_{\text{European election 2014}} \times \text{s.e.}_{\text{European election 2014}} + \text{s.e.}_{\text{municipal election 2013}} \times \text{s.e.}_{\text{municipal election 2013}}}$ OR $\text{s.e.}_{\text{national election 2015}} \times \text{s.e.}_{\text{national election 2015}})$ where h^2 indicates the narrow sense SNP heritability of turnout in each election and s.e. indicates the standard error of the estimate obtained from GCTA-GREML. A one-sided P value was further computed as $P = \text{pnorm}(-1 \times \text{abs}(z))$ using R. This method has previously been used to test the gender-specific differences in heritability estimates[94].

Genetic correlations. To obtain summary statistics to upload to the LD Hub[45], we performed GWASs of turnout in both samples for each of the three elections. The GWASs were performed using PLINK, adjusting for age, gender and the first ten principal components of genetic ancestry. LDSC h^2 estimates are provided in the Supplementary Appendix and Supplementary Table 11, and Manhattan and Q–Q plots are provided in Supplementary Fig. 4a–f). The loci which showed

association ($P < 1 \times 10^{-6}$) with electoral turnout are reported in Supplementary Table 10 by election for each sample. We computed genetic correlations of voter turnout with 72 traits categorized as smoking behaviour, neurological diseases, personality traits, reproductive traits, sleeping, cognitive, anthropometric traits, education, psychiatric diagnoses and aging on the LDHub. Significance thresholds were adjusted to account for multiple testing using Benjamini–Hochberg FDR correction[87], implemented using R.

Nagelkerke pseudo r^2. The variance explained in election turnout was calculated as the gain in pseudo r^2 after adding PS as the explanatory variable to a baseline logistic regression model with the voter turnout as the outcome and the first ten principal components of genetic ancestry, estimated using flashPCA[95], as covariates. Pseudo r^2 was calculated using the NagelkerkeR2 function in the fmsb R package (https://cran.r-project.org/package=fmsb). The bias-adjusted CIs were estimated using the R package boot[96,97] and 35,000 bootstrap replicates. The pseudo r^2 values were transformed to a liability scale using the population prevalence of voter turnout, turnout within each election and cohort using the equations in Lee et al.[98] (Supplementary Appendix 5.2; an analyses of robustness is provided in Supplementary Appendix 5.3).

Mendelian randomization analyses. We used the GCTA[70] with the summary statistics from the GWAS of voter turnout and summary statistics from the latest GWAS for educational attainment, intelligence test performance and height available in the public domain. Individuals of European origin in the 1000 Genome phase 3 dataset were utilized to supply linkage disequilibrium information between the SNPs. We conducted HEIDI outlier heterogeneity tests (further information about the Mendelian randomization analysis and robustness checks is provided in Supplementary Appendix 8 and 9).

One of the methodological assumptions with summary-statistics-based GSMR analysis is that there is no sample overlap between the GWAS samples utilized for the exposure and outcome of interest. We made sure that the iPSYCH samples used to perform GWASs of voter turnout were not used in meta-analysis of any of the exposure traits.

Another assumption of GSMR is that the effect of the exposure on the outcome is linear, which might not always be true. While we use the HEIDI test to detect and exclude pleiotropic SNPs, the power to detect pleiotropy relies on the sample size of the GWAS from which summary statistics are obtained and the effect sizes of the pleiotropic SNPs[99]. The large disparity in the sample sizes of the exposure and outcome traits restricts the elimination of all pleiotropic SNPs in our analysis.

Data analysis was not performed in a blinded manner.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Owing to the consent structure of the iPSYCH and Danish law, the data cannot be shared publicly owing to its sensitive nature. The data can be accessed from secure servers[59]. For further information, please contact P.B.M. (pbm@econ.au.dk). For access to the data in Supplementary Table 5, please contact K.M.H. (kmh@ifs.ku.dk) as these register data also cannot be shared publicly.

Code availability

Code and scripts available from the corresponding author on request.

Received: 30 August 2019; Accepted: 17 August 2020;
Published online: 09 November 2020

References

1. Gerber, A. S., Green, D. P. & Shachar, R. Voting may be habit-forming: evidence from a randomized field experiment. *Am. J. Polit. Sci.* **47**, 540–550 (2003).
2. Green, D. P. & Shachar, R. Habit formation and political behaviour: evidence of consuetude in voter turnout. *Br. J. Polit. Sci.* **30**, 561–573 (2000).
3. Plutzer, E. Becoming a habitual voter: inertia, resources, and growth in young adulthood. *Am. Polit. Sci. Rev.* **96**, 41–56 (2002).
4. Wolfinger, R. E. & Rosenstone, S. J. *Who Votes?* 22 (Yale Univ. Press, 1980).
5. Dawes, C. et al. The relationship between genes, psychological traits, and political participation. *Am. J. Polit. Sci.* **58**, 888–903 (2014).
6. Dawes, C. T., Settle, J. E., Loewen, P. J., McGue, M. & Iacono, W. G. Genes, psychological traits and civic engagement. *Philos. Trans. R. Soc. Lond. B* **370**, 20150015 (2015).
7. Fowler, J. H., Baker, L. A. & Dawes, C. T. Genetic variation in political participation. *Am. Polit. Sci. Rev.* **102**, 233–248 (2008).
8. Klemmensen, R. et al. The genetics of political participation, civic duty, and political efficacy across cultures: Denmark and the United States. *J. Theor. Polit.* **24**, 409–427 (2012).
9. Loewen, P. J. & Dawes, C. T. The heritability of duty and voter turnout. *Polit. Psychol.* **33**, 363–373 (2012).
10. Weinschenk, A. C., Dawes, C. T., Kandler, C., Bell, E. & Riemann, R. New evidence on the link between genes, psychological traits, and political engagement. *Polit. Life Sci.* **38**, 1–13 (2019).
11. Mondak, J. J. *Personality and the Foundations of Political Behavior* (Cambridge Univ. Press, 2010).
12. Mondak, J. J., Hibbing, M. V., Canache, D., Seligson, M. A. & Anderson, M. R. Personality and civic engagement: an integrative framework for the study of trait effects on political behavior. *Am. Polit. Sci. Rev.* **104**, 85–110 (2010).
13. Charney, E. Behavior genetics and postgenomics. *Behav. Brain Sci.* **35**, 331–358 (2012).
14. Charney, E. & English, W. Genopolitics and the science of genetics. *Am. Polit. Sci. Rev.* **107**, 382–395 (2013).
15. Frey, B. S. Why do high income people participate more in politics? *Publ. Choice* **11**, 101–105 (1971).
16. Tollison, R. D. & Willett, T. D. Some simple economics of voting and not voting. *Publ. Choice* **6**, 59–71 (1973).
17. Verba, S., Schlozman, K. L. & Brady, H. E. *Voice and Equality: Civic Voluntarism in American Politics* 4 (Harvard Univ. Press, 1995).
18. Brady, H. E., Verba, S. & Schlozman, K. L. Beyond SES: a resource model of political participation. *Am. Polit. Sci. Rev.* **89**, 271–294 (1995).
19. Hansen, K. M. Electoral turnouts: strong social norms of voting. in *Oxford Handbook of Danish Politics* (eds Christiansen, P. M. et al.) 76–87 (Oxford Univ. Press, 2020).
20. Bouchard, T. J. & McGue, M. Familial studies of intelligence: a review. *Science* **212**, 1055–1059 (1981).
21. Hill, W. D. et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol. Psychiatry* **24**, 169–181 (2018).
22. Plomin R., DeFries, J. C., Knopik, V. S. & Neiderhiser, J. M. *Behavioral Genetics* 6th edn (Worth Publishers, 2013).
23. Rowe, D. C., Jacobson, K. C. & Van den Oord, E. J. Genetic and environmental influences on vocabulary IQ: parental education level as moderator. *Child Dev.* **70**, 1151–1162 (1999).
24. Krapohl, E. et al. The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proc. Natl Acad. Sci. USA* **111**, 15273–15278 (2014).
25. Converse, P. E. In *The Human Meaning of Social Change* (eds Campbell, A. & Converse, P. E.) 263–337 (Russell Sage, 1972).
26. Dinesen, P. T. et al. Estimating the impact of education on political participation: evidence from monozygotic twins in the United States, Denmark and Sweden. *Polit. Behav.* **38**, 579–601 (2016).
27. Verba, S. & Nie, N. H. *Participation in America: Social Equality and Political Democracy* (Harper & Row, 1972).
28. Persson, M. Education and political participation. *Br. J. Polit. Sci.* **45**, 689–703 (2015).
29. Nie, N. H., Junn, J. & Stehlik-Barry, K. *Education and Democratic Citizenship in America* (Univ. Chicago Press, 1996).
30. Gerber, A. S. et al. Personality traits and participation in political processes. *J. Polit.* **73**, 692–706 (2011).
31. Richardson, M., Abraham, C. & Bond, R. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychol. Bull.* **138**, 353–387 (2012).
32. Moffitt, T. E. et al. A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl Acad. Sci. USA* **108**, 2693–2698 (2011).
33. Mõttus, R., Realo, A., Vainik, U., Allik, J. & Esko, T. Educational attainment and personality are genetically intertwined. *Psychol. Sci.* **28**, 1631–1639 (2017).
34. Belsky, D. W. et al. Genetic analysis of social-class mobility in five longitudinal studies. *Proc. Natl Acad. Sci. USA* **115**, E7275–E7284 (2018).
35. Belsky, D. W. et al. The genetics of success: how single-nucleotide polymorphisms associated with educational attainment relate to life-course development. *Psychol. Sci.* **27**, 957–972 (2016).
36. Barth, D., Papageorge, N. W. & Thom, K. Genetic endowments and wealth inequality. *J. Polit. Econ.* **128**, 1474–1522 (2020).
37. Sondheimer, R. M. & Green, D. P. Using experiments to estimate the effects of education on voter turnout. *Am. J. Polit. Sci.* **54**, 174–189 (2010).
38. Deary, I. J., Batty, G. D. & Gale, C. R. Childhood intelligence predicts voter turnout, voting preferences, and political involvement in adulthood: the 1970 British cohort study. *Intelligence* **36**, 548–555 (2008).
39. Hillygus, D. S. The missing link: exploring the relationship between higher education and political engagement. *Polit. Behav.* **27**, 25–47 (2005).
40. Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. The predictive value of IQ. *Merrill Palmer Q.* **47**, 1–41 (2001).
41. White, E. S. Intelligence and sense of political efficacy in children. *J. Polit.* **30**, 710–731 (1968).
42. White, K. R. The relation between socioeconomic status and academic achievement. *Psychol. Bull.* **91**, 461–481 (1982).

43. Strenze, T. Intelligence and socioeconomic success: a meta-analytic review of longitudinal research. *Intelligence* **35**, 401–426 (2007).
44. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
45. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
46. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
47. Johnson, W., McGue, M. & Iacono, W. G. Disruptive behavior and school grades: genetic and environmental relations in 11-year-olds. *J. Educ. Psychol.* **97**, 391–405 (2005).
48. Johnson, W., McGue, M. & Iacono, W. G. Genetic and environmental influences on academic achievement trajectories during adolescence. *Dev. Psychol.* **42**, 514–532 (2006).
49. Deary, I. J. & Johnson, W. Intelligence and education: causal perceptions drive analytic processes and therefore conclusions. *Int. J. Epidemiol.* **39**, 1362–1369 (2010).
50. Davies, G. et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
51. Davies, G. et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol. Psychiatry* **21**, 758–767 (2016).
52. Fowler, J. H. & Dawes, C. T. Two genes predict voter turnout. *J. Polit.* **70**, 579–594 (2008).
53. Abdellaoui, A. et al. Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* **3**, 1332–1342 (2019).
54. Ojeda, C. Depression and political participation. *Soc. Sci. Q.* **96**, 1226–1243 (2015).
55. Burden, B. C., Fletcher, J. M., Herd, P., Moynihan, D. P. & Jones, B. M. How different forms of health matter to political participation. *J. Polit.* **79**, 166–178 (2017).
56. Cesarini, D., Johannesson, M. & Oskarsson, S. Pre-birth factors, post-birth factors, and voting: evidence from Swedish adoption data. *Am. Polit. Sci. Rev.* **108**, 71–87 (2014).
57. Shultziner, D. Genes and politics: a new explanation and evaluation of twin study results and association studies in political science. *Polit. Anal.* **21**, 350–367 (2013).
58. Rosenstone, S. J. & Wolfinger, R. E. The effect of registration laws on voter turnout. *Am. Polit. Sci. Rev.* **72**, 22–45 (1978).
59. Pedersen, C. B. et al. The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
60. Highton, B. Voter identification laws and turnout in the United States. *Annu. Rev. Polit. Sci.* **20**, 149–167 (2017).
61. *Mental Disorders Affect One In Four People* World Health Report (WHO, 2001); https://www.who.int/whr/2001/media_centre/press_release/en/
62. Bullenkamp, J. & Voges, B. Voting preferences of outpatients with chronic mental illness in Germany. *Psychiatr. Serv.* **55**, 1440–1442 (2004).
63. Siddique, A. & Lee, A. A survey of voting practices in an acute psychiatric unit. *Ir. J. Psychol. Med.* **31**, 229–231 (2014).
64. Couture, J. & Breux, S. The differentiated effects of health on political participation. *Eur. J. Publ. Health* **27**, 599–604 (2017).
65. Ojeda, C. & Pacheco, J. Health and voting in young adulthood. *Br. J. Polit. Sci.* **49**, 1163–1186 (2017).
66. Sund, R., Lahtinen, H., Wass, H., Mattila, M. & Martikainen, P. How voter turnout varies between different chronic conditions? A population-based register study. *J. Epidemiol. Community Health* **71**, 475–479 (2017).
67. Kelly, B. D. & Nash, M. Voter participation among people attending mental health services in Ireland. *Ir. J. Med. Sci.* **188**, 925–929 (2018).
68. Evans, L. M. et al. Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity* **121**, 616–630 (2018).
69. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
70. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
71. Vinkhuyzen, A. A. et al. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Transl. Psychiatry* **2**, e102 (2012).
72. Benjamin, D. J. et al. The genetic architecture of economic and political preferences. *Proc. Natl Acad. Sci. USA* **109**, 8026–8031 (2012).
73. Tingsten, H. *Political Behavior: Studies in Election Statistics* (PS King, 1937).
74. Bhatti, Y., Dahlgaard, J. O., Hansen, J. H. & Hansen, K. M. Core and peripheral voters: predictors of turnout across three types of elections. *Polit. Stud.* **67**, 348–366 (2019).
75. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
76. Domingue, B. W. et al. The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health. *Proc. Natl Acad. Sci. USA* **115**, 702–707 (2018).
77. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9**, 96 (2017).
78. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
79. Blais, A. & St-Vincent, S. L. Personality traits, political attitudes and the propensity to vote. *Eur. J. Polit. Res.* **50**, 395–417 (2011).
80. Denny, K. & Doyle, O. Political interest, cognitive ability and personality: determinants of voter turnout in Britain. *Br. J. Polit. Sci.* **38**, 291–310 (2008).
81. Gallego, A. & Oberski, D. Personality and political participation: the mediation hypothesis. *Polit. Behav.* **34**, 425–451 (2012).
82. Weinschenk, A. Cause you've got personality: political participation and the tendency to join civic groups. *SAGE Open* **3**, 1–12 (2013).
83. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
84. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
85. Fowler, J. H. & Schreiber, D. Biology, politics, and the emerging science of human nature. *Science* **322**, 912–914 (2008).
86. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
87. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
88. Mortensen, P. B. Response to "Ethical concerns regarding Danish genetic research". *Mol. Psychiatry* **24**, 1574–1575 (2019).
89. Schork, A. J. et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
90. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish newborn screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
91. De Moor, M. H. et al. Meta-analysis of genome-wide association studies for personality. *Mol. Psychiatry* **17**, 337–349 (2012).
92. Bhatti, Y., Hansen, K. M. & Wass, H. The relationship between age and turnout: a roller-coaster ride. *Elect. Stud.* **31**, 588–593 (2012).
93. Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation* **3**, 139–141 (2008).
94. Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Correction: phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **14**, e1007228 (2018).
95. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS ONE* **9**, e93766 (2014).
96. Canty A. & Ripley B. D. boot: bootstrap R (S-Plus) functions. R package version 1.3-25 https://cran.r-project.org/web/packages/boot/index.html (2020).
97. Davison A. C. & Hinkley D. V. *Bootstrap Methods and Their Applications* (Cambridge Univ. Press, 1997); http://statwww.epfl.ch/davison/BMA/
98. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A. Better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* **36**, 214–224 (2012).
99. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).

Acknowledgements

We thank S. Oskarsson and members of the political behaviour research section at the Department of Political Science, Aarhus University for comments on previous versions of this manuscript. This research was supported by the Interacting Minds Centre, Aarhus University (seed grant no. 26223). The iPSYCH consortium is supported by the Lundbeck foundation (grant nos. R1-2=A9118 and R155-2014-1724). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

L.A., E.A. and M.B.P. conceived the study. L.A., V.A., A.J.S., A.B., E.A. and M.B.P. designed the study. L.A., V.A. and A.J.S. drafted the manuscript. L.A., V.A., A.J.S., K.M.H., E.A., A.B. and M.B.P. discussed the results and revised the manuscript. V.A. and A.J.S. conducted all of the analyses except that K.M.H. conducted the analyses for Supplementary Table 5. K.M.H. collected the turnout data. P.B.M., M.N., A.D.B., D.M.H., T.W., O.M. and E.A. designed, implemented, and/or oversaw the collection and generation of the IPSYCH data. All of the authors (L.A., V.A., K.M.H., A.J.S., T.W., O.M., A.D.B., D.M.H., M.N., P.B.M., W.K.T., A.B., E.A. and M.B.P.) approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00952-2.

Correspondence and requests for materials should be addressed to E.A.

Peer review information: Primary Handling Editor: Charlotte Payne.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Esben Agerbo

Last updated by author(s): Jul 8, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	All software used in data analysis are published in peer reviewed journals and referenced accordingly in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Because of the consent structure of the iPSYCH and Danish law, the data cannot be shared publicly due to its sensitive nature. The data can be accessed via secure servers (ref. 40: 9). For further information, please contact Professor Preben Bo Mortensen, Scientific Director of iPSYCH, (pbm@econ.au.dk).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Investigation of the genetic overlap between voter turnout and resources for politics, as indexed by educational attainment and intelligence test performance.
Research sample	We rely on the Danish iPSYCH case-cohort sample which consists of two large samples of comprehensively genotyped individuals: a nationally representative sample of individuals born in Denmark 1981–2005, and a population sample of all individuals born in the same period with one of the following psychiatric diagnoses: schizophrenia, mood disorders, bipolar affective disorder, autism, ADHD, and anorexia (see ref. 40 for details). The iPSYCH samples were merged with population-based records of whether Danish citizens voted in the 2015 national, 2014 European, and 2013 municipal elections in Denmark.
Sampling strategy	The nationally representative sample was collected using random sampling from the Danish CPR - where all people alive and living in Denmark are registered - of individuals born in Denmark 1981-2005. Using the CPR and the Danish Psychiatric Central Research Register a population sample of all individuals born 1981-2005 and diagnosed with schizophrenia, mood disorders, bipolar affective disorder, autism and attention-deficit/hyperactivity disorder was identified. These samples are state-of-the-field for single cohort GWAS studies, and constiitute the largest of its kind. The iPSYCH samples were merged with population-based records of whether Danish citizens voted in the 2015 national, 2014 European, and 2013 municipal elections in Denmark. We limit the analyze sample to individuals who were 18 years old or older - and hence had reach the Danish voting age to become eligible to vote in a public election in 2015.
Data collection	DNA was extracted from dried neonatal blood spots and amplified before genotyping on the Infinium PsychChip v1.0. Blood was collected between 4-7 days after birth and retrieved from the Danish Neonatal Screening Biobank within the Danish National Biobank. Psychiatric conditions were identified from national registers. Demographic and social variables were aggregated from national civil registers. Further details can be found in Ref. 40. Since 1997, the Danish Turnout Project has collected actual and validated turnout data from the 1,387 polling stations in Denmark. The turnout data covers about 70% of the population in each of the three elections. The municipality in which the individual polling station is location decides if voting files are made available to the Turnout Project, meaning that there is no individual selection. After the voting files were collected and verified, they were merged with individual social security number (CPR). See individual reports for each election, which describe this process in detail (refs. 2–4). In this case, the voting file with individual turnout data for the 2013 local, 2014 European, and 2015 national elections were merged into the iPSYCH data using the social security number (CPR).
Timing	All iPSYCH data was initially collected in 2012 and psychiatric diagnoses were later updated, complete through 2014. Turnout data were collected in 2103, 2014, and 2015 respectively (see above).
Data exclusions	Removed individuals who were not from a genetically homogeneous as population as identified by PCA analysis for association tests, heritability estimates. Removed individuals for whom voter turnout data was not available. We limit the analyze sample to individuals who were 18 years old or older - and hence had reach the Danish voting age to become eligible to vote in a public election in 2015.
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited according to consent by non-opt out from a national research program (see also the methods section in the main text and Schork et al 2019 in Nature Neuroscience 22, https://doi.org/10.1038/s41593-018-0320-0, and Mortensen 2018 in Molecular Genetics, https://doi.org/10.1038/s41380-018-0296-x, and Nørgaard-Pedersen & Hougaard 2007 in Journal of Inherited Metabolic Disease: Official Journal of the Society for the Study of Inborn Errors of Metabolism.
Ethics oversight	The iPSYCH has been approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, Statistics Denmark, and the Danish Neonatal Screening Biobank Steering Committee (see also the Methods section in the main text and Mortensen 2018 in Molecular Genetics, https://doi.org/10.1038/s41380-018-0296-x)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

nature research

https://doi.org/10.1038/s41562-020-00952-2

Supplementary information

Genetic predictors of educational attainment and intelligence test performance predict voter turnout

In the format provided by the authors and unedited

SI Appendix for

The genetics of citizenship: Genetic disposition towards educational attainment and intelligence test performance predict voter turnout

Table of Contents

S1. Supplementary details on the iPSYCH case-cohort sample and the turnout data

S1.1. Supplementary details on sampling for the iPSYCH Danish case-cohort sample

The study base for the iPSYCH Danish case-cohort sample consists of all singletons born with known mothers in Denmark between 1981 and 2005 as identified using the Danish CPR register (see ref. 1 for a detailed introduction to the sample). The iPSYCH case-cohort sample consists of a nationally representative sample drawn from the study base, and a psychiatric sample consisting of all individuals from the study base with a diagnosis of the following mental disorders: schizophrenia, mood disorders, bipolar affective disorder, autism, and attention-deficit/hyperactivity disorder. Since the iPSYCH2012 sample was initiated, "other related Danish projects have built on the same framework as that used within iPSYCH, for example, anorexia (5703 cases)" (ref. 1: 12). Individuals were genotyped sing neonatal dried blood spots from the Danish Neonatal Screening Biobank (Pedersen et al 2018: 7)

S1.1.1 Supplementary details on the genotyping process

DNA extraction was conducted at Statens Serum Institut in Denmark to prepare the samples for genotyping and sequencing. Genotyping was processed at the Broad Institute (Boston, MA, USA) "using the Infinium PsychChip v1.0 array (Illumina, San Diego, CA, USA) in accordance with the manufacturer's instructions" (Pedersen et al 2018: 9). Quality control was performed, and genotyped samples with call rates below 95% were rejected (N = 2270) (Pedersen et al 2018: 9).

S1.1.2 Supplementary details on Imputation and Quality Control

The iPSYCH imputation and data quality control has previously been extensively documented by e.g. Schork et al 2019, and the section below draws extensively on this article.

The iPSYCH data includes 78,050 samples that were genotyped at 554,360 markers in 23 waves. The 1000 genomes phase 3 call set in VCF format was downloaded for 414 samples belonging to the Yoruba Indian (YRI), Han Chinese (CHB), JPT (Japanese in Tokyo) and Central European ancestry from Utah (CEU) populations.

SNPs with a minor allele frequency less than 5% within each population, SNPs violating hardy Weinberg equilibrium ($p > 10^{-6}$) and SNPs not present in iPSYCH call set, and SNPs with pairwise $r^2 > 0.1$ within a 1kb region were excluded, which resulted in 101,532 SNPs.

Principal component analysis was performed using Eigensoft (https://www.hsph.harvard.edu/alkes-price/software/) software using the resulting SNPs and the iPSYCH samples were projected into this principal component space. Outlier detection algorithm ABERRANT (http://www.well.ox.ac.uk/software) was used with an inlier to outlier standard deviation ratio of 1:20 in the first two principal components to identify 75,501 samples of a homogenous genetic origin to perform SNP QC.

SNPs with a minor allele frequency less than 1%, violating hardy Weinberg equilibrium ($p < 10^{-6}$) within the subset of control samples of the previously defined individuals of homogenous genetic origin were excluded. After excluding NDELs, multiallelic loci and non-autosomal, strand ambiguous SNPs, a total of 246,539 SNPs were phased using SHAPEIT3 and further imputed to 80,707,375 SNPs in 10 batches using IMPUTE2 and the 1000 genomes phase 3 dataset as the reference.

After imputation, 17,324,340 SNPs were excluded for low INFO scores (< 0.2), 67,580,963 SNPs were excluded for low minor allele frequency (< 0.001), 1,831,455 SNPs were censored for

high missing rate (> 10% of samples). A further 291,937 SNPs were removed for significant associations to one of the 23 genotyping waves ($5x10^{-8}$). 33 SNPs were excluded for significant associations with imputation batch ($p < 5x10^{-8}$). 527,912 SNPs were excluded owing to differential imputation quality between cases and controls, 954 more SNPs were excluded for deviations from hardy Weinberg equilibrium in controls ($p > 1x10^{-6}$).

As part of sample QC, 22 samples were flagged for abnormal heterozygosity that could not be explained by admixture or runs of homozygosity, 283 samples were excluded for high missingness (> 1% of genotyped SNPs), 13 samples were excluded for discordance between annotated gender in the national registers and expected heterozygosity from X chromosome markers. 48 samples were excluded for duplication, in each case retaining the sample with lower missing genotype rate.

A strategy that combined traditional principal component analysis with Danish nationwide birth registers was used to select a set of homogenous European individuals for complex trait analyses. Initially the 1000 genomes phase 3 dataset was subset to include the set of high quality SNPs that were used in the iPSYCH pre-phasing and imputations. Strand ambiguous A/T and G/C SNPs, SNPs with minor allele frequency < 5% and missingness > 2.5% were excluded, followed by selection of independent SNPs using PLINK's indep-pairwise module. The MHC region was also excluded, which resulted in 43,789 SNPs for ethnicity QC. Eigensoft smartPCA module was used to compute principal components using the 1000 genomes samples and the iPSYCH samples were projected onto this principal component space.

Using the Danish national birth records, the mean and standard deviations were computed for the first 10 principal components of 47,586 individuals with both parents and both sets of grandparents born in Denmark. Subsequently, for each of the iPSYCH samples, the Mahalanobis distance from the multivariate mean of the joint distribution of the first 10 principal components from the 47,586 individuals of Danish origin was computed and assuming a chi-square distribution with 10 degrees of freedom, 6,474 samples whose distance had a probability < $5x10^{-7}$ were flagged as global ethnicity outliers.

A second outlier removal step involved computing PCs after pruning the SNPs for independence in the iPSYCH samples that pass the global ethnicity QC. Using the same Mahalanobis distance, a further 689 samples were excluded as local ancestry outliers.

This resulted in a sample set of 65,535 subjects and 11,601,089 SNPs that could be used for complex trait association analyses (Schork et al 2019).

S1.2. Supplementary information on the data from the Danish Turnout Project

Since 1997, the Danish Turnout Project has collected actual and validated turnout data from the 1,387 polling stations in Denmark. The turnout data covers about 70% of the population in each of the three elections analyzed in this paper. The municipality in which the individual polling station is located decides if voting files are made available to the Turnout Project, meaning that there is no individual selection. After the voting files were collected and verified, they were merged with individual social security number (CPR). See individual reports for each election, which describe this process in detail (Bhatti et al 2014a, 2014b, 2016). In this case, the voting file with individual turnout data for the 2013 local, 2014 European, and 2015 national elections were merged into the iPSYCH data using the social security number (CPR).

Specifically, in the nationally representative sample in the iPSYCH, voter turnout information was available for 7,746 individuals for the European parliament election in 2014, 11,110 individuals for the national election in 2015, and for 12,475 individuals for the municipal elections in 2013. In the psychiatric sample, voter turnout information was available for 18,071 individuals for the European election in 2014, 25,920 individuals for the national election in 2015, and for 25,920 individuals for the municipal election in 2013.

S2. Summary statistics and correlational statistics

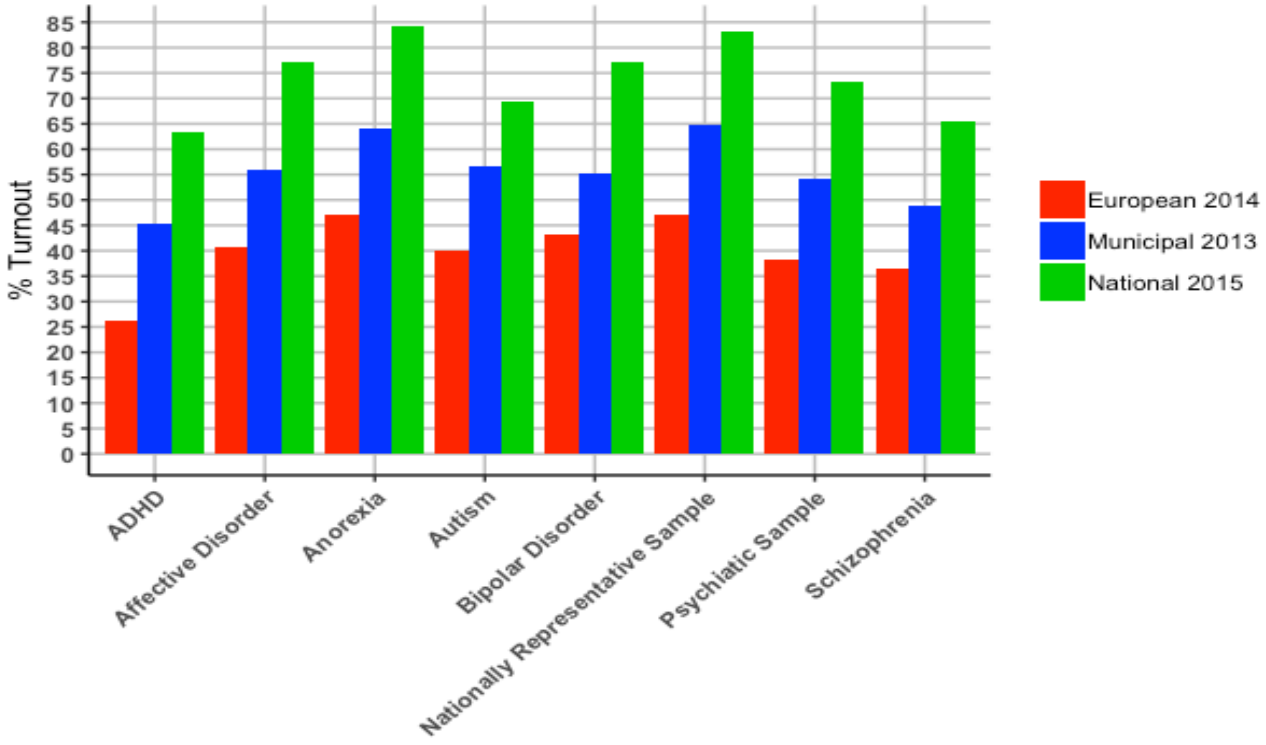
S2.1 The proportion of voters and non-voters by sample, election, and psychiatric condition

Supplementary Table S1 describes the proportion of voters and non-voters in the three elections in the psychiatric sample and the nationally representative sample. As a baseline for comparison, the table also reports the proportion of voters and non-voters in the Danish population in each of the three elections. Testifying to the high representativeness of the nationally representative sample and the measurement validity of our observed data on actual turnout, there is a relatively high match between the proportion of non-voters in the nationally representative sample and the population result for each election. Supplementary Figure 1 summarizes electoral turnout by psychiatric condition and for the overall psychiatric and nationally representative sample.

Supplementary Table 1. Proportion voters and non-voters in each election by sample and for the Danish population

Election	Sample	N	Non-voting (%)	Voting (%)
2015 National election	Psychiatric	21044	27	73
	Nationally representative	10225	17	83
	Danish population result		*14*	*86*
2014 European election	Psychiatric	14347	62	38
	Nationally representative	6818	53	47
	Danish population result		*44*	*56*
2013 Municipal election	Psychiatric	19255	46	54
	Nationally representative	9091	35	65
	Danish population result		*28*	*72*

Supplementary Figure 1. Turnout in % by psychiatric condition and for the overall psychiatric sample and the nationally representative sample



S2.2 Summary statistics for voters and non-voters

SupplementaryTables S2–S4 describe the characteristics of voters and non-voters on polygenic scores (PS) for education and intelligence test performance as well as demographic background variables for each of the three elections in the nationally representative sample and the psychiatric sample.

Supplementary Table 2. Summary statistics for voters and non-voters, 2015 national election

	Nat. rep sample		Psychiatric sample	
	Voters	**Non-voters**	**Voters**	**Non-voters**
Gender (proportion female)	0.5085	0.4127	0.5769	0.4181
Age in days at election day (mean)	10857	10616	10039	9817
PS for intelligence test performance (mean)	-2.26E-05	-2.29E-05	-2.27E-05	-2.30E-05
PS for educational attainment (mean)	6.30E-05	6.25E-05	6.28E-05	6.25E-05
PS for Openness (mean)	0.001	0.001	0.001	0.001
PS for Extraversion (mean)	0.0539	0.0537	00537	0.0538
PS for Neuroticism (mean)	0.0924	0.0925	0.0924	0.0924
PS for Conscientiousness (mean)	-0.0018	-0.0018	-0.0018	-0.0018
PS for Agreeableness (mean)	0.0008	0.0008	-0.0008	0.0008
Proportion with psychiatric a condition	0.0424	0.0837	1	1
n	9247	1863	18963	6955

Note. Entries are means except for the information on gender distribution and psychiatric condition where we report proportions.

Supplementary Table 3. Summary statistics for voters and non-voters, 2014 European election

	Nat. rep sample		Psychiatric sample	
	Voters	**Non-voters**	**Voters**	**Non-voters**
Gender (proportion female)	0.5316	0.4678	0.5893	0.5247
Age in days at election day (mean)	10790	10582	10246	10029
PS for intelligence test performance (mean)	-2.33E-05	-2.25E-05	-2.25E-05	-2.30E-05
PS for educational attainment (mean)	6.31E-05	6.30E-05	6.30E-05	6.26E-05
PS for Openness (mean)	0.0009	0.001	0.0009	0.001
PS for Extraversion (mean)	0.054	0.0538	0.0537	0.0537
PS for Neuroticism (mean)	0.0924	0.0925	0.0924	0.0923
PS for Conscientiousness (mean)	-0.0018	-0.0018	-0.0018	-0.0018
PS for Agreeableness (mean)	-0.0008	-0.0008	-0.0008	-0.0008
Proportion with a psychiatric condition	0.0986	0.1467	1	1
n	3638	4108	6906	11163

Note. Entries are means except for the information on gender distribution and psychiatric condition where we report proportions.

Supplementary Table 4. Summary statistics for voters and non-voters, 2013 Municipal election

	Nat. rep sample		Psychiatric sample	
	Voters	**Non-voters**	**Voters**	**Non-voters**
Gender (proportion female)	0.5169	0.4493	0.5864	0.4852
Age in days at election day (mean)	10440	10382	10210	10167
PS for intelligence test performance (mean)	-2.25E-05	-2.29E-05	-2.26E-05	-2.30E-05
PS for educational attainment (mean)	6.30E-05	6.26E-05	6.29E-05	6.25E-05
PS for Openness (mean)	0.001	0.001	0.001	0.001
PS for Extraversion (mean)	0.0538	0.0538	0.0537	0.0537
PS for Neuroticism (mean)	0.0924	0.0924	0.0924	0.0924
PS for Conscientiousness (mean)	-0.0018	-0.0018	-0.0018	-0.0018
PS for Agreeableness (mean)	-0.0008	-0.0008	-0.0008	-0.0008
Proportion with a psychiatric condition	0.101	0.1678	1	1
n	8084	4391	16391	13899

Note. Entries are means except for the information on gender distribution and psychiatric condition where we report proportions.

S2.3. Correlations between turnout and educational achievement and cognitive performance in Denmark

Through Statistics Denmark, the full data set from the Danish Turnout project has been linked to individual-level administrative data on educational attainment and cognitive performance as measured by the grade point in the cross-national written exam in mathematics from the 9th form-level school-leaving examination of the Danish compulsory schooling.

The written examination in mathematics in a given year is mandatory for all students across Denmark when they finish the 9th form of the Danish folkeskole. We are able to link to grade point data from exam years 2000 to 2014 to the voter files. That provides us with 256,151 (municipal election), 158,269 (European election) and 209,636 (national election) individuals for whom we have both turnout data and grade point data available. We have information about both educational attainment and turnout available for about 2-4 million individuals depending on the election.

Because of legal restrictions it is not possible to link the individual-level administrative data on educational attainment and grade point in mathematics into the iPSYCH data at the time where the current research is conducted. Therefore, we cannot use these variables in the analyses of the iPSYCH data. Yet, we can use the validated register data to investigate the correlation between turnout and educational attainment and cognitive performance, respectively, predicted by the resource model of political participation (see the introduction in the main text).

Supplementary Table 5 below shows the gamma correlation between turnout and educational attainment, and cognitive performance as measured by grade points in mathematics from the 9th form-level school-leaving examination of the Danish compulsory schooling. Turnout is coded 0-1, were ”1” = voted, and ”0” = did not vote. Educational attainment is measured in five categories: 1) primary or lower secondary school, 2) vocational training, 3) high school, 4) short post-secondary education, 5) graduate degree. Grade point in written mathematics is measured on a scale ranging from ”0” to ”13” which was the official grading scale in Denmark in 2000, where “13” was the highest grade.

The results in Supplementary Table 5 show moderate to strong correlations between voter turnout and educational attainment and cognitive performance as measured by grade point in mathematics. These results are consistent with past international research on the correlation between turnout and educational attainment and cognitive performance, respectively. We also find moderate correlations between educational attainment and cognitive performance as indexed by the grade point in written mathematics from the school-leaving examination.

Supplementary Table 5. Correlation between turnout and educational attainment and cognitive performance, respectively

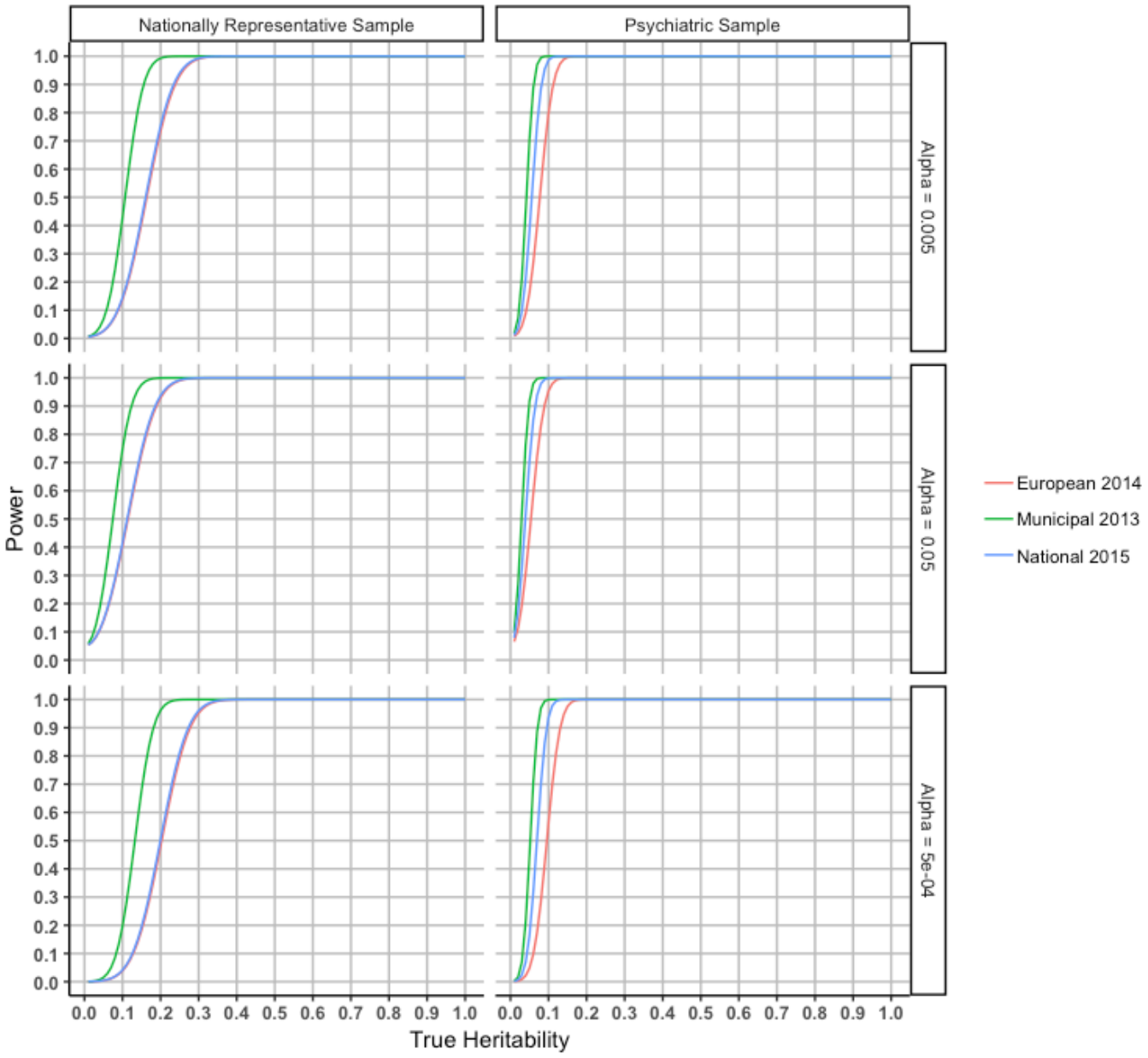
Correlation between turnout and educational attainment			
Election	**Gamma correlation**	**ASE**	**N**
2013 municipal election[1]	0.294	0.001	4,168,713
2014 European election[1]	0.333	0.001	2,292,683
2015 national election[1]	0.405	0.001	3,043,148
Correlation between turnout and grade point in mathematics			
2013 municipal election	0.287	0,003	256,151
2014 European election	0.347	0.003	158,269
2015 national election	0.343	0.004	209,636
Correlation between educational attainment and grade point in mathematics			
2013 municipal election	0.230	0,003	256,151
2014 European election	0.280	0.003	158,826
2015 national election	0.330	0.002	209,636

Note. For further information about the turnout data, see Bhatti et al. (2016, 2014a, 2014b).

[1] In the subsample for whom we have data on grade point in mathematics and turnout behavior, the gamma correlation between educational attainment and voter turnout is for municipal, 0.274, ASE 0.002, N=256,171, European 0.230, ASE 0.004, N=158,269, National 0.3296, ASE 0.002, N = 209.636.

S3. Power analysis

Supplementary Figure 2. Power to reject the null hypothesis using GCTA-GREML estimates of narrow sense heritability explained by common SNPs for different values of true heritability and type I error with the sample sizes available in each sample and election



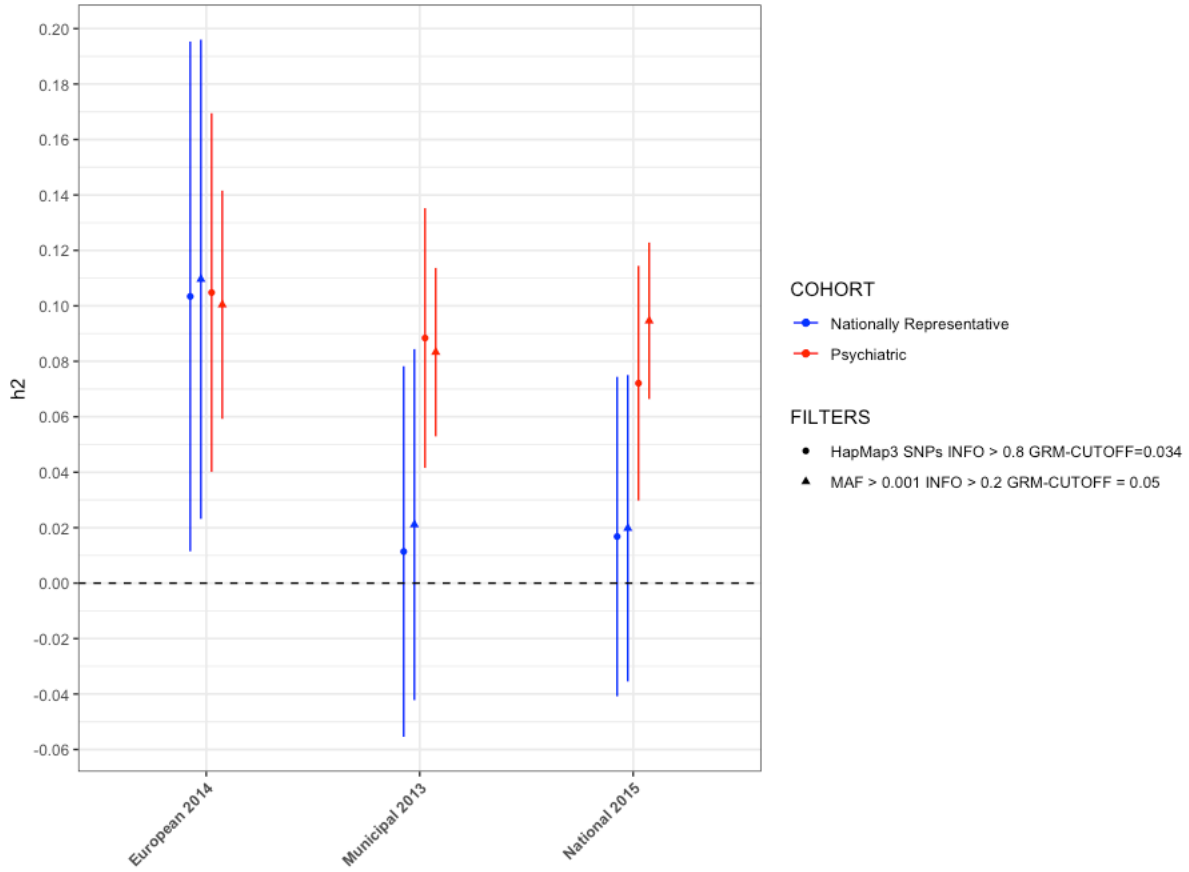
Note. The lines in Supplementary Figure 2 illustrate the power to reject the null hypothesis using GCTA-GREML estimates of narrow sense heritability explained by common SNPs for different values of true heritability and alpha levels of 0.005 (upper panels), 0.05 (panels in the middle) and 5e-04 (lower panels) with the sample sizes available in the nationally representative sample and the psychiatric sample for each election.

S4. Supplemental information for the analyses of the heritability of turnout

S4.1. Analysis of robustness for Figure 1 in the main text

As we describe in the methods section in the main text, "the GRM for estimating the GCTA-GREML SNP heritability was computed using the standard SNP filters chosen for iPSYCH data release (MAF > 0.001, Imputation INFO score > 0.2) and we used a grm-cutoff of 0.05 as a threshold for relatedness between samples." To check the robustness of the results reported in Figure 1 in the main text, we conducted an analysis of robustness where we build a GRM from SNPs common between the HapMap3 dataset (Duan et al 2008) and iPSYCH with minor allele frequency ≥ 0.01 and imputation INFO scores ≥ 0.8 and a more stringent relatedness threshold with a grm-cutoff of 0.034. The results are reported in Supplementary Figure 3 below which compares the estimates obtained with the standard SNP filters chosen for iPSYCH data and the more stringent filter. As seen in Supplementary Figure 3, the point estimates of SNP heritability only change very slightly and importantly, the statistical significance of our results does not change.

Supplementary Figure 3. Comparison of heritability estimates for voting in municipal, European, and national elections by sample for SNP filters with minor allele frequency > 0.001 and imputation INFO score > 0.2 against a stringent SNP filtering criteria of minor allele frequency ≥ 0.01 and imputation INFO score ≥ 0.8. 95% confidence intervals.



S4.2. Family-based heritability estimates

Family-based estimates of heritability are useful for providing estimates of the total contribution of genetic factors to a phenotype of interest. We used tetrachoric correlations (Pearson 1900, Olsson 1979) to describe the resemblance between individuals in one of four familial pairings: spousal, mother–offspring, father–offspring, and sibling. In Supplementary Tables 5.1–2 below, we report the spousal resemblance and estimates of heritability on the liability scale, accounting for spousal resemblance (Falconer & Mackay 1996) for each genetic relative type as well as an inverse variance weighted meta-analysis (Wray & Gottesman 2012) combining the three estimates.

S4.2.1 Methods

For each individual ascertained for the iPSYCH case-cohort study (Pedersen 2011) (N = 30,000 ascertained randomly; N = 58,764 ascertained for psychiatric outcomes), we used the Danish Civil Registration System (15) to collect all possible familial relatives through mother–father–child links (N = 1,930,812 unique individuals). SOLAR v7.6.4 (Almasy et al 1998) was used to define all of the pairwise pedigree relationships. We defined two sets of relatives for analysis: a random cohort set and a case cohort set. The random cohort set sampled from pedigrees seeded with the 30,000 randomly ascertained iPSYCH individuals to ensure they represent random draws from the Danish population and are not enriched for psychiatric outcomes. The case cohort set samples relative pairs from the entire combined pedigree set where both individuals had a registered diagnosis of a psychiatric condition. Within each set of relative pairs, any one individual is only observed once and pairs were discarded until there were no annotated familial relationships across pairs. Only those with complete voting data were retained.

Spousal resemblance was estimated via tetrachoric correlation implemented in the *polychor* (Olsson 1979) package for R. P-values are approximate, computed using a two-sided z-test assuming the sampling distribution of the tetrachoric correlation coefficient is normal with a standard deviation equal to the estimated standard error.

A general method for estimating heritability (Falconer & Mackay 1996) under the assumption of no shared environmental effects is to estimate the phenotypic correlation of traits measured in a chosen type of relatives and divide this by their expected kinship coefficient (i.e., 0.5 for 1st degree relatives). For dichotomous traits, the tetrachoric correlation coefficient (Pearson 1900) can be used to estimate a phenotypic correlation for an underlying latent bivariate liability. Similarly, it can also be used to estimate heritability, as long as the 2×2 table of dichotomous phenotypes is represented of the sampled population. Spousal resemblance (i.e., correlations in the phenotypes of parents) changes the expectations of the phenotypic correlations and adjustments are needed to avoid based estimates of heritability (Falconer & Mackay 1996). We use the following equations, taken from (Falconer & Mackay 1996), where r_{Tet} is the tetrachoric correlation estimated among different relative types (PO, Parent Offspring; sib, Full Sibling; Spouse, parents).

$$h^2_{PO} = (\frac{1}{0.5} r_{Tet,PO})/(1 + r_{Tet,Spouse})$$

$$h^2_{sib} = (\frac{1}{0.5} r_{Tet,sib})/(1 + r_{Tet,Spouse} h^2_{PO})$$

The estimator of heritability for siblings in the context of spousal resemblance is dependent on the heritability itself. We use the inverse variance weighted combination of the two parental estimates in place of the true value. We again use approximate p-values compute from a two-sided z-test. We meta-analyzed all three estimates using inverse variance weighting.

S4.2.2 Results

Supplementary Tables 6.1–2 show the spousal resemblance and estimates of heritability on the liability scale, accounting for spousal resemblance (Falconer & Mackay 1996) for each genetic relative type as well as an inverse variance weighted meta-analysis (Wray & Gottesman 2012) combining the three estimates.

The findings show a strong resemblance among spouses for electoral participation (r_{Tet} ranging from 0.39 to 0.78), suggesting assortative mating or shared environmental causes (Falconer & Mackay 1996). The estimates of heritability are consistently moderate, ranging from 0.39 to 0.49 across the election cycles, and highly significant ($p < 1.24 \times 10^{-40}$). Our approach for heritability estimation assumes the contributions of common environment to be negligible, which may not be true given the strong resemblance among spouses. These estimates should therefore be seen as upper limits for our population.

Supplementary Table 6.1. Spousal resemblance and estimates of heritability on the liability scale, accounting for spousal resemblance. Nationally representative sample.

Election	Relationship	N (pairs)	N00	N01	N10	N11	r_{Tet}	r_{Tet} s.e.	h^2	h^2 s.e.	p (Approximate)
Municipal 2013	Spousal	25791	2988	2358	2884	17561	0.64	0.01	-	-	0.00E+00
	Mother–offspring	17187	2281	4416	1345	9145	0.43	0.01	0.48	0.02	1.62E-127
	Father–offspring	16475	2164	4233	1483	8595	0.38	0.01	0.42	0.02	5.15E-100
	Sibling	4998	581	788	810	2819	0.33	0.02	0.49	0.05	5.55E-23
	Meta-analysis								**0.46**	**0.01**	**1.07E-245**
European 2014	Spousal	11620	3327	1240	1138	5915	0.78	0.01	-	-	0.00E+00
	Mother–offspring	8502	2623	2205	804	2870	0.51	0.01	0.53	0.03	1.84E-69
	Father–offspring	7818	2339	2116	813	2550	0.45	0.02	0.46	0.03	2.55E-53
	Sibling	1855	524	351	371	609	0.34	0.03	0.47	0.07	2.88E-11
	Meta-analysis								**0.49**	**0.02**	**6.17E-130**
National 2015	Spousal	16266	610	894	1087	13675	0.59	0.01	-	-	0.00E+00
	Mother–offspring	13089	673	2013	758	9645	0.44	0.02	0.49	0.03	2.36E-60
	Father–offspring	12228	547	1958	827	8896	0.34	0.02	0.38	0.04	2.16E-21
	Sibling	3104	165	330	340	2269	0.38	0.03	0.56	0.07	1.21E-15
	Meta-analysis								**0.46**	**0.02**	**2.17E-92**

Note. For each election, we use family data in the form of relative pairs aggregated from the Danish civil registers a family-based heritability. Families were sampled from genealogies seeded with a member of the iPSYCH random sample. Individual estimates from mother-offspring, father-offspring, and sibling pairs were adjusted for expected spousal correlations and meta-analyzed using inverse variance weighting. N00, number of non-voting pairs (both relatives); N01, number of pairs with relative 1 voting and relative 2 non-voting; N10, number of pairs with relative 1 non-voting and relative 2 voting; N11, number of voting pairs (both relatives); r_{Tet}, tetrachoric correlation of voting in relative pairs; r_{Tet}, s.e., standard error of tetrachoric correlation; h^2, heritability estimate; h^2 s.e., standard error of heritability estimate; p, p-value for the significance of the heritability estimate.

Supplementary Table 6.2. Spousal resemblance and estimates of heritability on the liability scale, accounting for spousal resemblance. Psychiatric sample

Election	Relationship	N (pairs)	N00	N01	N10	N11	rTet	rTet s.e.	h2	h2 s.e.	p (Approximate)
Municipal 2013	Spousal	3271	874	566	608	1223	0.42	0.02	-	-	6.56E-98
	Mother–offspring	9171	2205	2476	1196	3294	0.33	0.02	0.42	0.03	9.41E-44
	Father–offspring	6193	1517	1686	916	2074	0.27	0.02	0.34	0.04	5.07E-17
	Sibling	2292	610	479	488	715	0.24	0.03	0.39	0.06	9.68E-11
	Meta-analysis								**0.39**	**0.02**	**1.20E-67**
European 2014	Spousal	1390	602	198	223	367	0.56	0.03	-	-	9.24E-78
	Mother–offspring	4293	1784	1047	491	971	0.44	0.02	0.47	0.04	3.05E-32
	Father–offspring	2668	1105	637	343	583	0.4	0.03	0.43	0.06	8.32E-13
	Sibling	979	440	155	195	189	0.37	0.05	0.53	0.09	2.98E-09
	Meta-analysis								**0.47**	**0.03**	**6.76E-51**
National 2015	Spousal	2023	230	251	351	1191	0.39	0.04	-	-	1.84E-22
	Mother–offspring	6756	677	1465	676	3938	0.34	0.02	0.43	0.04	2.28E-26
	Father–offspring	4237	439	980	546	2272	0.22	0.03	0.28	0.05	1.86E-08
	Sibling	1719	213	267	308	931	0.32	0.04	0.52	0.08	1.13E-10
	Meta-analysis								**0.39**	**0.03**	**1.24E-40**

Note. For each election, we use family data in the form of relative pairs aggregated from the Danish civil registers a family-based heritability. Families were sampled from genealogies seeded with a member of the iPSYCH case sample. Individual estimates from mother-offspring, father-offspring, and sibling pairs were adjusted for expected spousal correlations and meta-analyzed using inverse variance weighting. N00, number of non-voting pairs (both relatives); N01, number of pairs with relative 1 voting and relative 2 non-voting; N10, number of pairs with relative 1 non-voting and relative 2 voting; N11, number of voting pairs (both relatives); r_{Tet}, tetrachoric correlation of voting in relative pairs; r_{Tet}, s.e., standard error of tetrachoric correlation; h^2, heritability estimate; h^2 s.e., standard error of heritability estimate; p, p-value for the significance of the heritability estimate.

S5. Supplementary information about the polygenic score analyses

S5.1 Supplementary details on the base datasets for the polygenic scores and the target dataset

The base dataset for the educational attainment scores was based on summary statistics from 10,101,243 SNPs (670,392 after clumping and intersection with target dataset) and 766,345 individuals obtained from the social science genetic association consortium website (https://www.thessgac.org/data) as described in Lee et al (2018).

The base dataset for the intelligence test performance scores was based on summary statistics from 9,295,119 SNPs (491,784 after clumping and intersection with target dataset) and 269,867 individuals as described in Savage et. al 2018 and downloaded from https://ctg.cncr.nl/software/summary_statistics.

The base datasets for the big five personality traits was based on summary statistics from 175,375 adult individuals, obtained from the genetics of personality consortium (http://www.tweelingenregister.org/GPC/) as described in de Moor et al (2012).

The target dataset was the best guess genotypes from subsets of iPSYCH individuals for whom turnout data was available for each election.

S5.2. Nagelkerke pseudo r^2 estimates

Supplementary Table 7 below shows the Nagelkerke pseudo r^2 estimates for the variance in electoral turnout that can be accounted for by the polygenic scores for educational attainment and intelligence test performance, respectively, in the psychiatric and the nationally representative sample.

Supplementary Table 8 shows the Nagelkerke pseudo r^2 estimates for the variance in electoral turnout that can be accounted for by the PS for each of the big five personality traits.

Specifically, the r^2 estimates in Supplemetary Tables S7 and S8 represent the increment in Nagelkerke pseudo r^2-estimates when polygenic score was added to a baseline logistic regression model with voter turnout as the outcome and the first ten principal components of genetic ancestry as explanatory variables. p-values are significance of polygenic score from logistic regressions.

Supplementary Table 7. Variance in turnout accounted for by polygenic scores for educational attainment and intelligence test performance

Election	Sample	PS for educational attainment	PS for intelligence test performance
Municipal 2013	Nationally Representative	r^2 = 0.0134 (95% CI: 0.0091 – 0.0186); r^2_L = 0.0162; p = 2.92e^{-28}	r^2 = 0.0058 (95% CI: 0.0032 – 0.0094); r^2_L = 0.007; p = 3.09e^{-13}
	Psychiatric	r^2 = 0.0156 (95% CI: 0.0126, 0.019); r^2_L = 0.0194; p < 2.47e^{-78}	r^2 = 0.0069 (95% CI: 0.005 – 0.0093); r^2_L = 0.0086; p = 4.38e^{-36}
European 2014	Nationally Representative	r^2 = 0.0317 (95% CI: 0.0235 – 0.0413); r^2_L = 0.0374; p = 3.83e^{-41}	r^2 = 0.01 (95% CI: 0.0056 – 0.0158); r^2_L = 0.012; p = 2.52e^{-14}
	Psychiatric	r^2 = 0.0278 (95% CI: 0.0225 – 0.0337); r^2_L = 0.032; p = 7.71e^{-81}	r^2 = 0.01 (95% CI: 0.0077 – 0.0148); r^2_L = 0.012; p = 2.39e^{-33}
National 2015	Nationally Representative	r^2 = 0.0199 (95% CI: 0.0139 – 0.0273); r^2_L = 0.0263; p = 3.01e^{-30}	r^2 = 0.0041 (95% CI: 0.0016 – 0.0078); r^2_L = 0.0054; p = 1.61^{e-7}
	Psychiatric	r^2 = 0.0125 (95% CI: 0.0095 – 0.016); r^2_L = 0.0189; p < 3.48e^{-50}	r^2 = 0.0049 (95% CI: 0.0031 – 0.0071); r^2_L = 0.0074; p = 1.01e^{-20}

Note. r^2 = The increment in Nagelkerke pseudo r^2-estimates when polygenic score was added to a baseline logistic regression model with voter turnout as the outcome and the first ten principal components of genetic ancestry as explanatory variables. p-values are significance of polygenic score from logistic regressions. r^2_L = r^2 transformed to a liability scale using population prevalence of voter turnout and case proportion within each sample and election and the equations from Lee et. al 2012 (DOI: 10.1002/gepi.21614). Pseudo r^2-estimates were calculated using NagelkerkeR2 function as implemented in the R package fmsb version 0.6.3 (https://www.rdocumentation.org/packages/fmsb/versions/0.6.3). 95% bias adjusted confidence intervals were obtained by ordinary non-parametric bootstrap, implemented using the R package boot (Canty 2020, Davidson 1997) and 35,000 bootstrap replicates.

Supplementary Table 8. Variance in turnout accounted for by polygenic scores for the big 5 personality traits

Election	Sample	PS for agreeableness	PS for openness	PS for neuroticism	PS for extra-version	PS for conscien-tiousness
Municipal 2013	Nationally Representative	r^2 = 0.0014; r^2_L = 0.0017; p = 0.00027	r^2 = 0.0006; r^2_L = 0.0007; p = 0.0161	r^2 = 0.0005; r^2_L = 0.0006; p = 0.034	r^2 = 8.7e^{-6}; r^2_L = 1.04e^{-5}; p = 0.7783	r^2 = 5.6e^{-6}; r^2_L = 6.74e^{-6}; p = 0.8213
	Psychiatric	r^2 = 0.0007; r^2_L = 0.0009; p = 4.32e^{-05}	r^2 = 6.47e^{-5}; r^2_L = 8.07e^{-5}; p = 0.2254	r^2 = 0.0002; r2_L = 0.0003; p = 0.0252	r^2 = 1.03e^{-5}; r^2_L = 1.28e^{-5}; p = 0.6282	r^2 = 9E-09; r^2_L = 1.12e^{-08}; p = 0.9885
European 2014	Nationally Representative	r^2 = 0.0015; r^2_L = 0.0017; p= 0.00305	r^2 = 0.0014; r^2_L = 0.0016; p = 0.0044	r^2 = 2e^{-10}; r^2_L = 2.34e^{-10}; p = 0.99	r^2 = 1.62^{e-7}; r^2_L = 1.92e^{-7} p = 0.9754	r^2 = 1.04e^{-6}; r^2_L = 1.23e^{-6}; p = 0.938
	Psychiatric	r^2 = 0.0007; r^2_L = 0.0008; p = 0.0014	r^2 = 0.0007; r^2_L = 0.0008; p = 0.0016	r^2 = 5.45^{e-5}; r^2_L = 6.28^{e-5}; p = 0.3945	r^2 = 0.0003; r^2_L = 0.0004; p = 0.0404	r^2 = 1.65e^{-5}; r2_L = 1.9e^{-5}; p = 0.6393
National 2015	Nationally Representative	r^2 = 0.0006; r^2_L = 0.0008; p = 0.0431	r^2 = 1.09e^{-5}; r^2_L = 1.44e^{-05}; p = 0.78	r^2 = 0.001; r^2_L = 0.0013; p = 0.0095	r^2 = 2.61e^{-5}; r^2_L = 3.45e^{-5}; p=0.6773	r^2 = 0.0002; r^2_L = 0.0003; p = 0.2116
	Psychiatric	r^2 = 0.0005; r^2_L = 0.0009; p = 0.00128	r^2 = 0.00036; r^2_L = 0.00054; p = 0.0112	r^2 = 0.0001; r^2_L = 0.0002; p = 0.0974	r^2 = 6.07e^{-5}; r^2_L = 9.19e^{-5}; p = 0.2977	r^2 = 8.23e^{-5}; r2_L = 0.00012; p = 0.2254

Note. r^2 = The increment in Nagelkerke pseudo r^2-estimates when polygenic score was added to a baseline logistic regression model with voter turnout as the outcome and the first ten principal components of genetic ancestry as explanatory variables. p-values are significance of polygenic score from logistic regressions. r^2_L = r^2 transformed to a liability scale using population prevalence of voter turnout and case proportion within each sample and election and the equations from Lee et. al 2012 (DOI: 10.1002/gepi.21614). Pseudo r^2-estimates were calculated using NagelkerkeR2 function as implemented in the R package fmsb version 0.6.3 (https://www.rdocumentation.org/packages/fmsb/versions/0.6.3). 95% bias adjusted confidence intervals were obtained by ordinary non-parametric bootstrap, implemented using the R package boot (Canty 2020, Davidson 1997) and 35,000 bootstrap replicates.

S5.3 Analyses of robustness

For the polygenic score calculations, we chose the default iPSYCH QC parameters to ensure a good overlap of SNPs between the reference and target summary statistics. To check the robustness of the results reported in Supplementary Table 6, we repeated the analysis with a more stringent SNP selection criteria of minor allele frequency ≥ 0.01 and imputation INFO score ≥ 0.8. The results are reported in Supplementary Table 9 which provides a comparison of the polygenic score prediction performance of educational attainment and intelligence test performance on each election within each sample for SNP filters with minor allele frequency > 0.001 and imputation INFO score > 0.2 against a stringent SNP filtering criteria of minor allele frequency > 0.01 and imputation INFO score > 0.8. The results in Supplementary Table 9 show that while the variance explained is slightly different, the significance of the results do not change.

Supplementary Table 9. Comparison of the variance explained in turnout by polygenic scores for educational attainment and intelligence test performance by election and sample for SNP filters with minor allele frequency > 0.001 and imputation INFO score > 0.2 against a stringent SNP filtering criteria of minor allele frequency > 0.01 and imputation INFO score > 0.8

Election	Sample	PS for educational attainment		PS for intelligence test performance	
		MAF > 0.001; INFO > 0.2	MAF ≥ 0.01; INFO ≥ 0.8	MAF > 0.001; INFO > 0.2	MAF ≥ 0.01; INFO ≥ 0.8
Municipal 2013	Nationally Representative	r^2 = 0.0134; p = $2.92e^{-28}$	r^2 = 0.0169; p = $4.17e^{-35}$	r^2 = 0.0058; p = $3.09e^{-13}$	r^2 = 0.005; p = $1.1e^{-11}$
	Psychiatric	r^2 = 0.0156; p = $2.47e^{-78}$	r^2 = 0.0194; p = $9.27e^{-97}$	r^2 = 0.0069; p = $4.38e^{-36}$	r^2 = 0.0038; p = $6.53e^{-21}$
European 2014	Nationally Representative	r^2 = 0.0317; p = $3.83e^{-41}$	r^2 = 0.0371; p = $1.35e^{-47}$	r^2 = 0.01; p = $2.52e^{-14}$	r^2 = 0.0049; p = $8.52e^{-08}$
	Psychiatric	r^2 = 0.0278; p = $7.71e^{-81}$	r^2 = 0.032; p = $1.59e^{-92}$	r^2 = 0.01; p = $2.39e^{-33}$	r^2 = 0.008; p = $6.28e^{-25}$
National 2015	Nationally Representative	r^2 = 0.0199; p = $3.01e^{-30}$	r^2 = 0.0228; p < $2.04e^{-34}$	r^2 = 0.0041; p = $1.61e^{-07}$	r^2 = 0.0023; p = $7.8e^{-05}$
	Psychiatric	r^2 = 0.0125; p = $3.48e^{-50}$	r^2 = 0.0149; p < $2.81e^{-59}$	r^2 = 0.0049; p = $1.01e^{-20}$	r^2 = 0.0027; p = $3.02e^{-12}$

Note. r^2 = The increment in Nagelkerke pseudo r^2-estimates when polygenic score was added to a baseline logistic regression model with voter turnout as the outcome and the first ten principal components of genetic ancestry as explanatory variables. p-values are significance of polygenic score from logistic regressions. Pseudo r^2-estimates were calculated using NagelkerkeR2 function as implemented in the R package fmsb version 0.6.3 (https://www.rdocumentation.org/packages/fmsb/versions/0.6.3). 95% bias adjusted confidence intervals were obtained by ordinary non-parametric bootstrap, implemented using the R package boot (Canty 2020, Davidson 1997) and 35,000 bootstrap replicates. MAF refers to "minor allele frequency" and INFO refers to imputation info scores.

S6. Supplementary information about the genome wide association analysis

S6.1. Methods

Additive genotype dosages were obtained from the genotype probabilities emitted by Impute2 for the 65,535 subjects and 11,601,089 SNPs that pass our quality control criterion as described in supplementary information S1.1.2.

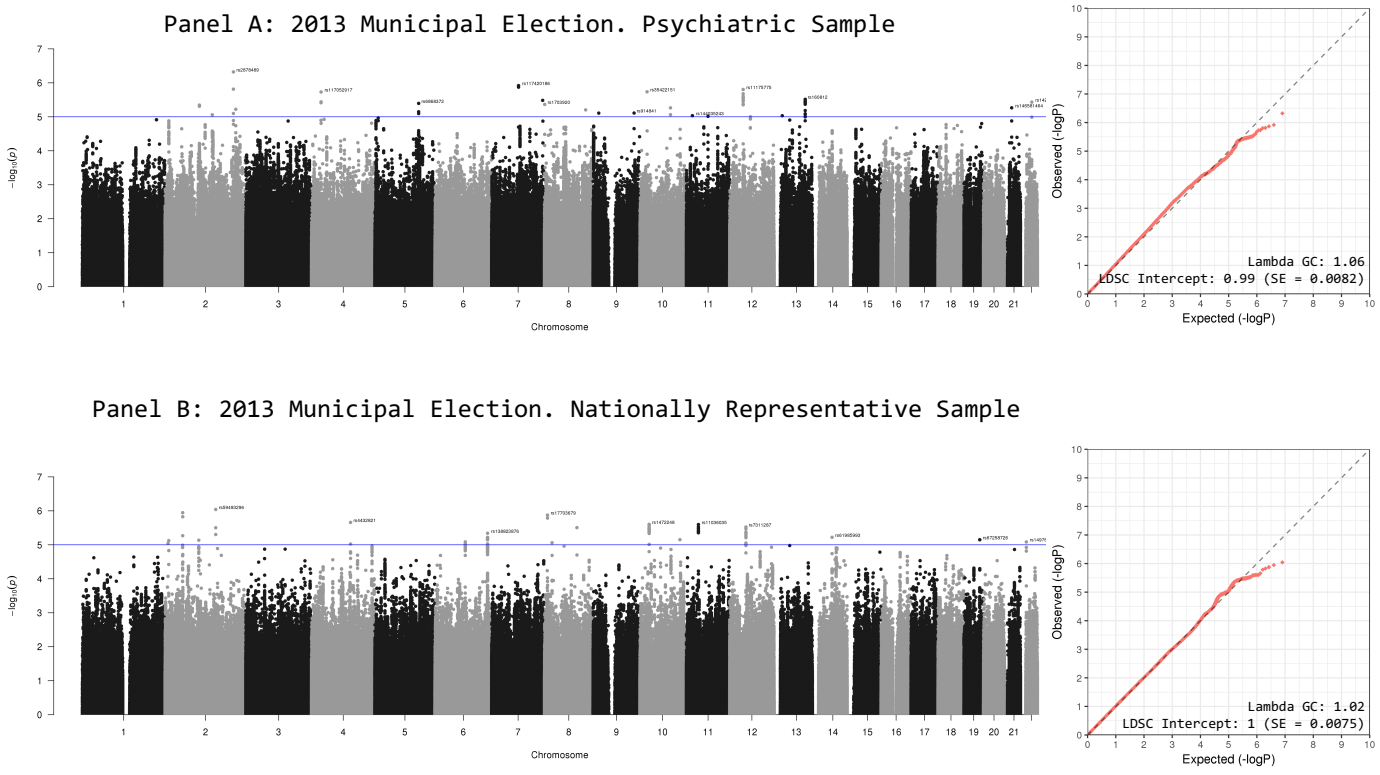
As described in S2, in the nationally representative sample, voter turnout information was available for 7,746 individuals for the European parliamentary election in 2014, 11,110 individuals for the national election in 2015 and 12,475 individuals for the municipal election in 2013. In the psychiatric sample, voter turnout information was available for 18,071 individuals for the European election in 2014, 25,920 individuals for the national election in 2015 and 25,920 individuals for the municipal election in 2013.

Genome wide association studies were performed using the logistic function in PLINK with whether or not an individual voted in a specific election as the binary outcome, the additive genotype dosages as explanatory variables and age, age-squared, gender and the first ten principal components of genetic ancestry as covariates.

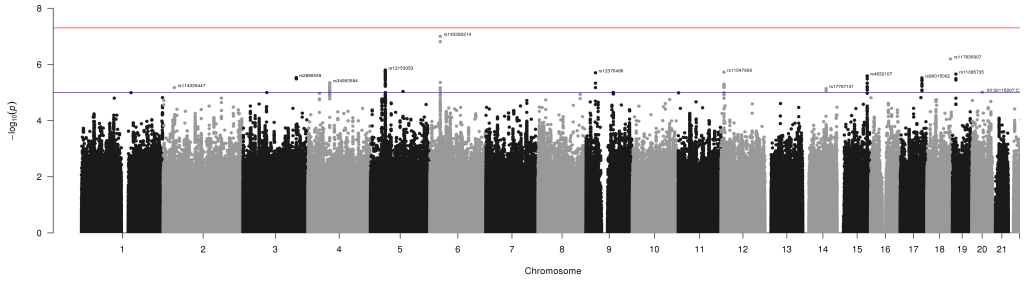
S6.2. Results

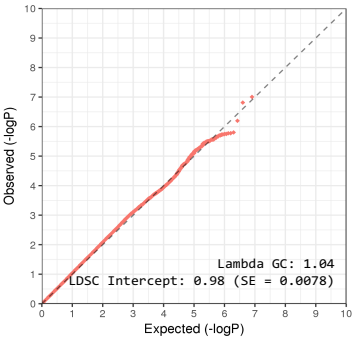
Supplementary Figure 4, panels a-f, show the Manhattan and quantile quantile-plots. The loci that show association ($p < 1x10^{-6}$) with electoral turnout in each of the three elections in our two samples are shown in Supplementary Table 10 below.

Supplementary Figure 4. Manhattan and Q-Q plots

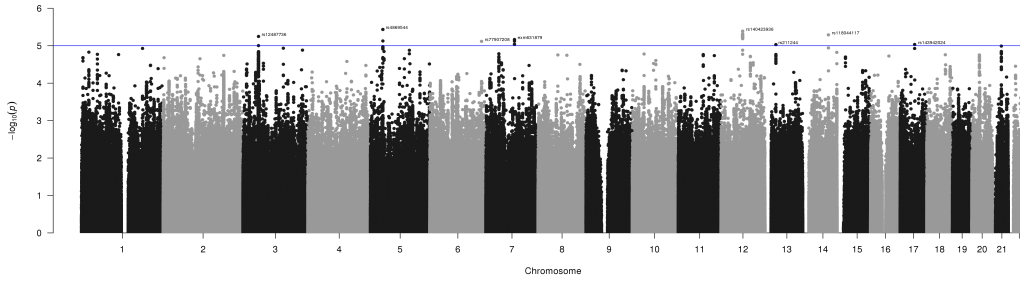


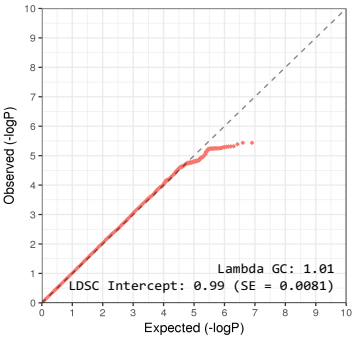
Panel C: 2014 European Election. Psychiatric Sample



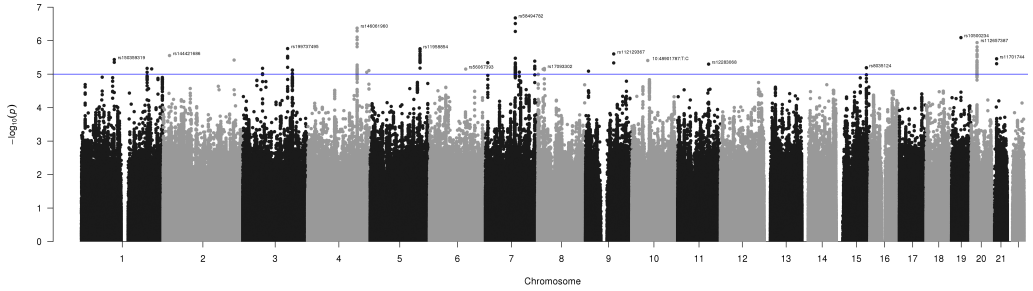


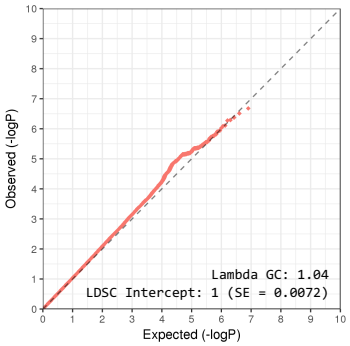
Panel D: 2014 European Election. Nationally Representative Sample



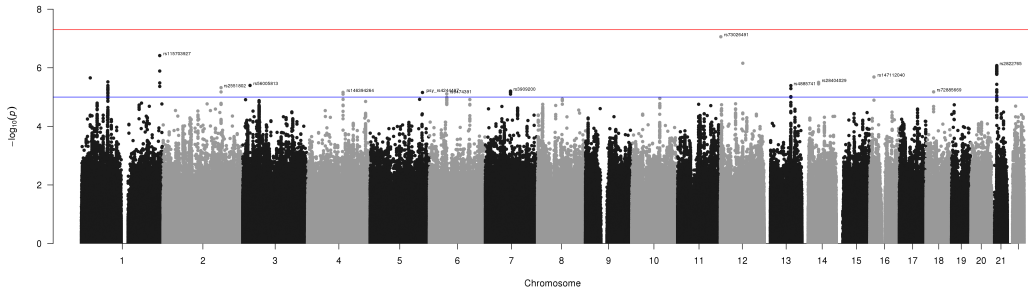


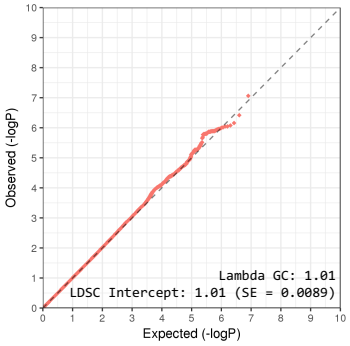
Panel E: 2015 National Election. Psychiatric Sample





Panel F: 2015 National Election. Nationally Representative Sample





Supplementary Table 10. Loci that show association ($p < 1x10^{-6}$) with electoral turnout in each of the three elections in our two samples

CHR	BP	SNP	A1	A2	FRQ	INFO	OR[1]	SE	p	Election	Sample
4	151175987	rs4235252	T	A	0.9468	0.9686	1.2388	0.0437	9.38E-07	National 2015	Psychiatric
4	151196494	rs146061960	T	TA	0.9458	1.0027	1.24	0.0425	4.22E-07	National 2015	Psychiatric
4	151202952	rs11099759	T	C	0.9454	1.0029	1.2332	0.0424	7.85E-07	National 2015	Psychiatric
4	151209530	rs6842890	T	C	0.947	1.0013	1.241	0.043	5.12E-07	National 2015	Psychiatric
7	91190283	rs58494782	A	AT	0.1273	0.9103	0.8519	0.0309	2.10E-07	National 2015	Psychiatric
7	91192292	rs2430448	A	G	0.1141	0.9371	0.8522	0.0319	5.30E-07	National 2015	Psychiatric
7	91195485	rs2160256	G	A	0.1152	0.9397	0.8503	0.0317	3.08E-07	National 2015	Psychiatric
19	28676721	rs10500234	G	C	0.1314	0.9762	0.8648	0.0295	8.11E-07	National 2015	Psychiatric
6	31883679	exm-rs9267673	C	T	0.0786	1.0025	0.8044	0.0415	1.54E-07	European 2014	Psychiatric
6	31899598	rs143383214	C	T	0.0783	0.9993	0.801	0.0416	9.93E-08	European 2014	Psychiatric
18	72074509	rs117635907	C	T	0.0104	0.4312	0.4071	0.1804	6.34E-07	European 2014	Psychiatric
2	205841191	rs2878489	G	A	0.653	0.9772	0.9156	0.0175	4.76E-07	Municipal 2013	Psychiatric
1	239033711	rs115703927	G	T	0.0135	0.7711	0.4683	0.1494	3.83E-07	National 2015	Nationally Representative
12	1523459	rs73026491	T	C	0.0192	0.7374	0.4974	0.1305	8.71E-08	National 2015	Nationally Representative
12	68570702	rs78491746	A	G	0.0389	0.7694	0.6229	0.0954	7.00E-07	National 2015	Nationally Representative
21	15915723	rs2822760	C	T	0.4632	0.9888	0.8358	0.0365	9.01E-07	National 2015	Nationally Representative
21	15918714	rs12329657	A	G	0.4624	0.9882	0.8362	0.0365	9.70E-07	National 2015	Nationally Representative
21	15921516	rs2822765	A	G	0.4623	0.9924	0.8357	0.0364	8.40E-07	National 2015	Nationally Representative
21	15925156	rs2178934	A	T	0.4632	0.9963	0.8364	0.0364	9.04E-07	National 2015	Nationally Representative
2	152497544	rs59493296	T	C	0.0769	0.988	0.7846	0.0494	9.09E-07	Municipal 2013	Nationally Representative

Note. CHR represents the chromosome and BP refers to the base position within the chromosome in accordance with the human genome reference version hg19. A1 and A2 are the two loci observed at each locus. FRQ refers to the frequency of A2 within the analysis population. INFO score is a metric of the uncertainty in the missing data imputation at each locus. [1]OR: Odds ratios from logistic regression with respect to A2

S7. Supplementary information about genetic correlations using LDHub

The summary statistics from the six genome wide association studies described in S6 were intersected with the SNP list recommend by LDHub (de Moor et al 2012), and odds ratios were converted to Z-scores in R using the equation Z = abs(qnorm(P-value/2)). The resulting summary statistics were uploaded to LDHub and the traits categorized as smoking behaviour, anthropometric traits, neurological diseases, personality traits, reproductive traits, education, blood lipids, brain volume, glycaemic traits, psychiatric diseases, cardio-metabolic traits, ageing and sleeping were selected for computing genetic correlations.

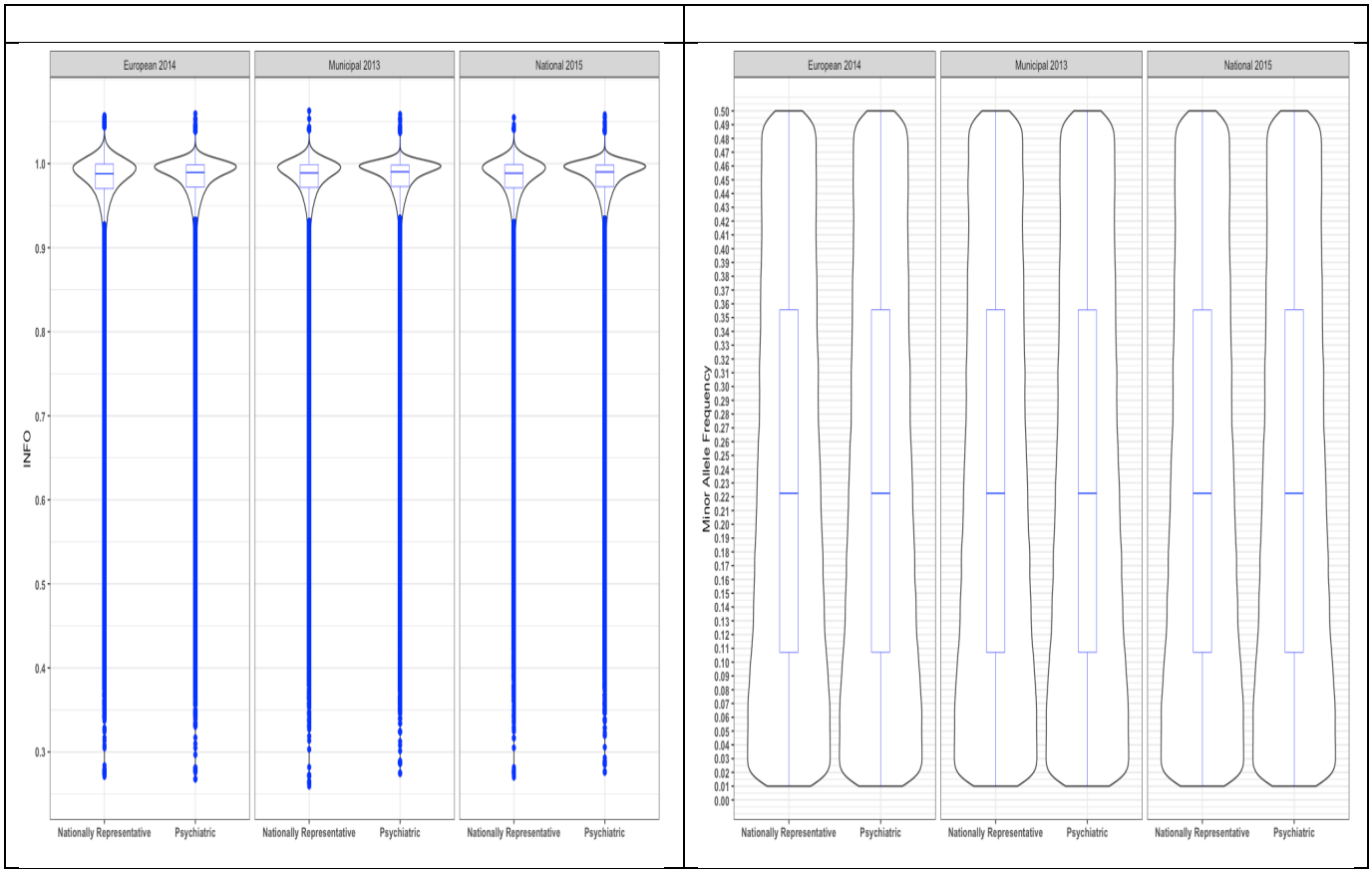
Supplementary Table 11. LDSC h2 estimates by election and sample

Election	Sample	LDSC h2 (Observed Scale)	SE
National 2015	Psychiatric Sample	0.0869	0.0199
National 2015	Nationally Representative Sample	0.0115	0.0495
European 2014	Psychiatric Sample	0.1602	0.0295
European 2014	Nationally Representative Sample	0.1499	0.07
Municipal 2013	Psychiatric Sample	0.1056	0.0169
Municipal 2013	Nationally Representative Sample	0.0474	0.0364

S7.1. Analyses of robustness

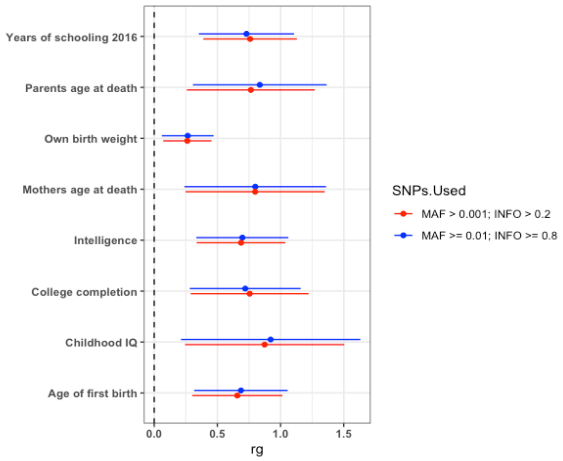
For the LDHub analysis, we use a set of 971,720 markers that are common between the iPSYCH dataset and the list of markers supplied by LDHub (http://ldsc.broadinstitute.org/static/media/w_hm3.noMHC.snplist.zip). All of the SNPs in this subset had a minor allele frequency ≥ 0.01 in iPSYCH and only a small minority were imputed with an INFO score < 0.8.

Supplementary Figure 5: Violin plot of the imputation INFO scores and minor allele frequencies of the iPSYCH SNPs and their associations with voter turnout used to compute genetic correlations using LDHub.



To further verify the robustness of our findings, we applied more stringent filters to select SNPs with a minor allele frequency ≥ 0.01 and imputation INFO score ≥ 0.8. Using these filters, we checked the genetic correlations that were found to be significant for the 2014 European election in the nationally representative sample. As seen in Supplementary Figure 6, we find that the results and significance remain unchanged.

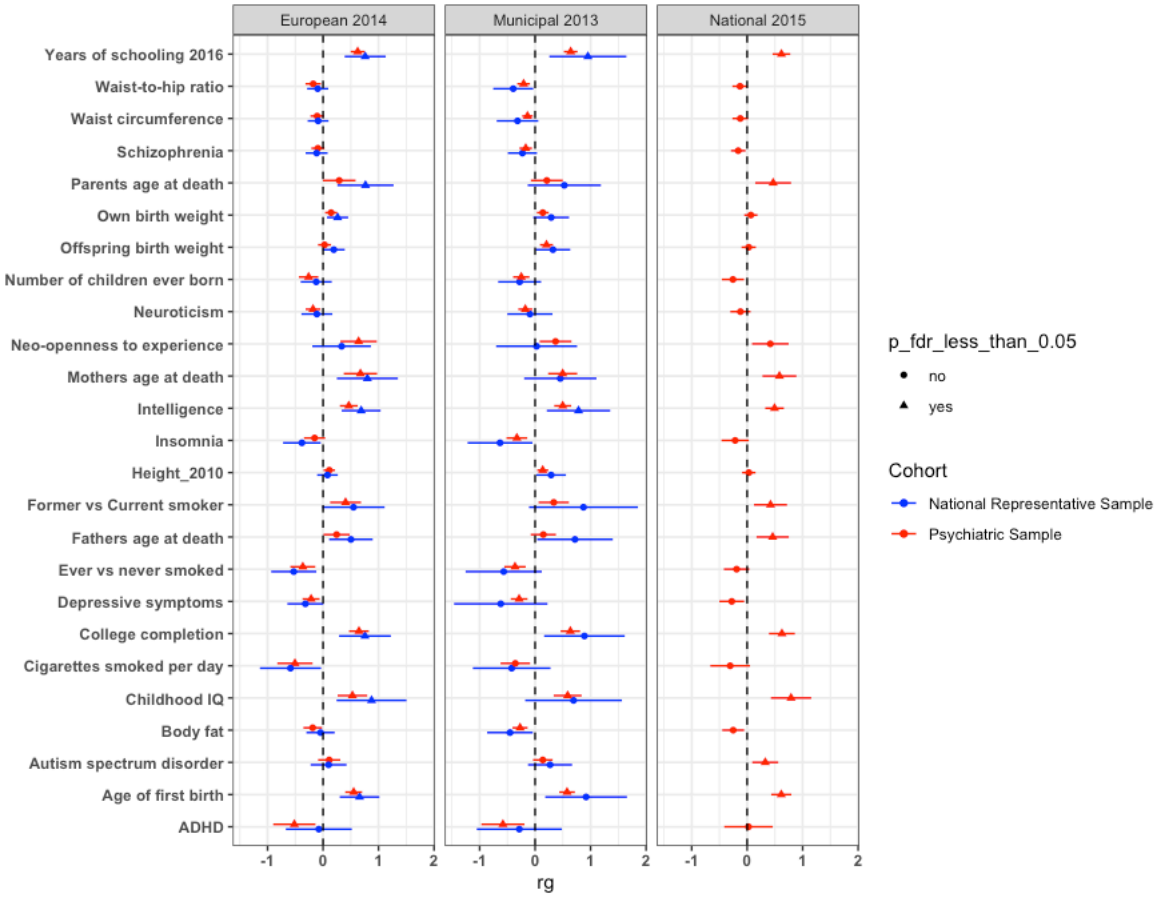
Supplementary Figure 6. Analysis of robustness of the genetic correlations under more stringent SNP selection between voter turnout and non-voting traits that show significant genetic correlation after FDR adjustment in the nationally representative sample for the 2014 European election. 95% confidence intervals.



S7.2. Supplemental analyses for the interpretation of Figure 3 in the main text

As noted in the main text, the results reported in Figure 3 show positive genetic correlations for electoral turnout across all three elections with indicators of educational attainment (i.e., college completion and years of schooling) and intelligence test performance (i.e., childhood IQ). As we also note in the main text when discussing the results in Figure 3, as expected, the correlations in the nationally representative sample follow the same pattern, albeit with higher p-values and larger error bars owing most likely to the smaller sample size. As we describe in the main text, "none of the genetic correlations in Figure 3 show any discernible difference between the two samples for any of the three elections. These results can be seen in the results reported in Supplementary Figure 7. It reports the same results as Figure 3 in the main text but organized in a way that maximizes possibility to compare estimates in the nationally representative sample and the psychiatric sample. Specifically, by election, Supplementary Figure 7 presents the genetic correlations between voter turnout and non-voting traits that show significant genetic correlation after FDR adjustment in at least one of the turnout phenotypes in the two samples. 95% confidence intervals.

Supplementary Figure 7. The genetic correlations between voter turnout and non-voting traits that show significant genetic correlation after FDR adjustment in at least one of the turnout phenotypes in the two samples organized by election. 95% confidence intervals.



Note. The LDSC estimated heritability of turnout in the 2015 national election was not significantly different from zero in the nationally representative sample; hence, no correlations were computed.

S8. Supplementary information about the summary statistics based Mendelian randomization analysis

S8.1. Methods

The summary statistics based mendelian randomization was performed using the GSMR module of the GCTA software suit version 1.92.1 beta 6. The summary statistics for educational attainment and intelligence were obtained as described in S5. The summary statistics for height were based on summary statistics from 253,288 individuals across 2,511,830 markers as described in Wood et al (2014).

The genome wide significance threshold for SNP inclusion was set at $p < 5x10^{-8}$, LD $r^2 < 0.05$ and the HEIDI threshold for removal of horizontal pleiotropic SNPs was set at 0.01. For the summary statistics based mendelian randomization analyses using GSMR with educational attainment, intelligence test performance and height as exposure variables, we choose the genome wide significant SNPs in each of these genome wide association studies, which are by default in the common allele frequency spectrum and imputed with high confidence in European populations. A further check of the data confirmed this as none of the SNPs have a minor allele frequency < 0.01 or an imputation INFO score < 0.6. The LD reference was generated from the European individuals of the 1000 genomes phase 3 dataset, which was downloaded in VCF format and converted to PLINK binary file format using PLINK.

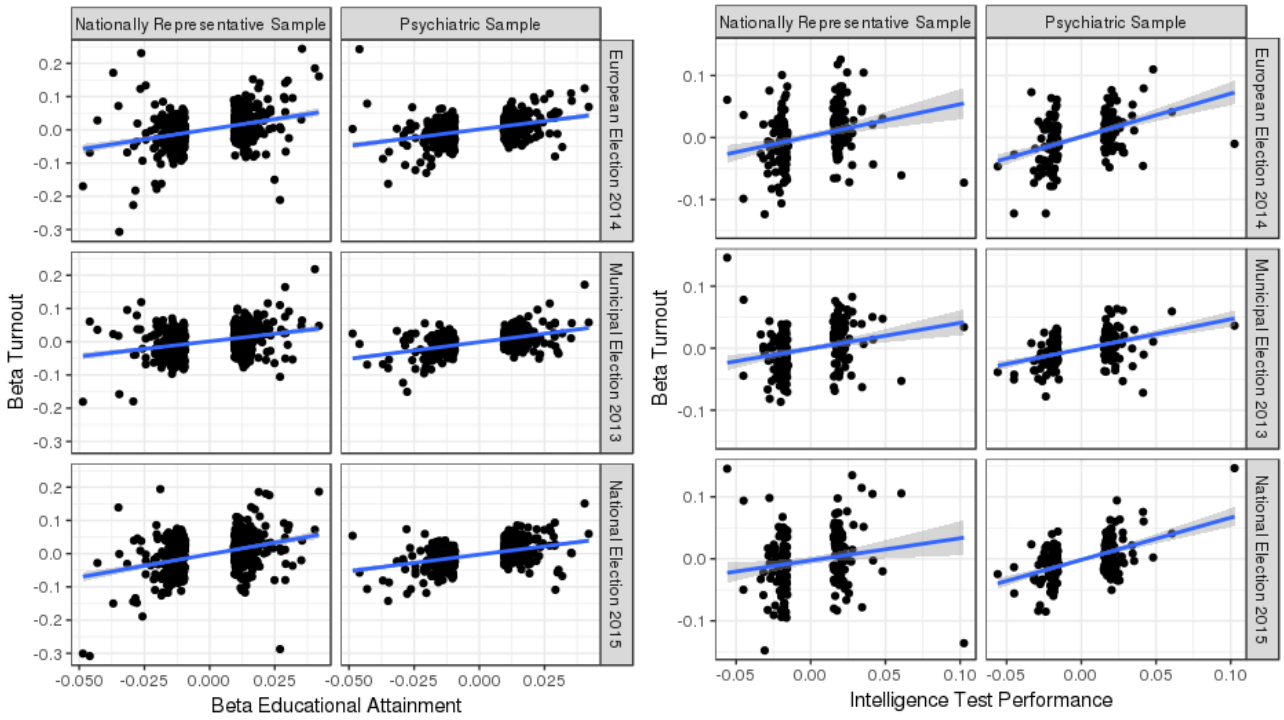
S8.2. Supplementary results from the Mendelian randomization analysis

Supplementary Table 12. Results from the Mendelian randomization analysis

Sample	Exposure	Turnout outcome	b_{xy}	OR	SE	p-value	nSNPs	HEIDI Outlier SNPs
Psychiatric	Educational attainment	2015 national election	0.974835	2.65	0.0800013	3.72E-34	512	1
	Intelligence test performance	2015 national election	0.577715	1.78	0.0789845	2.58E-13	211	0
	Height	2015 national election	0.0452416	1.04	0.0322309	0.16	744	4
	Educational attainment	2014 European election	1.09113	2.97	0.0866865 1	3.44E-36	512	1
	Intelligence test performance	2014 European election	0.729359	2.07	0.0858192	1.91E-17	210	1
	Height	2014 European election	0.0662612	1.06	0.0349562	0.058	744	3
	Educational attainment	2013 municipal election	0.983106	2.67	0.0656989	1.26E-50	512	1
	Intelligence test performance	2013 municipal election	0.510974	1.66	0.064776	3.06E-15	211	0
	Height	2013 municipal election	0.0930759	1.09	0.0253689	0.0004	747	1
Nationally Repre-sentative	Educational attainment	2015 national election	1.20255	3.32	0.143968	6.66E-17	512	1
	Intelligence test performance	2015 national election	0.61527	1.85	0.142014	1.47E-05	211	0
	Height	2015 national election	0.176583	1.19	0.0580355	0.0023	745	3
	Educational attainment	2014 European election	1.14436	3.14	0.129059	7.51E-19	512	1
	Intelligence test performance	2014 European election	0.767272	2.15	0.127706	1.87E-09	210	1
	Height	2014 European election	-0.044649	0.95	0.0520331	0.39	747	1
	Educational attainment	2013 municipal election	0.978357	2.66	0.106406	3.76E-20	513	0
	Intelligence test performance	2013 municipal election	0.574856	1.77	0.10517	4.6E-08	211	0
	Height	2013 municipal election	0.0289206	1.02	0.0429828	0.5	747	1

Note. b_{xy} refers to the effect size of the exposure on outcome. SE, p-value are respectively the standard error, associated p-value of the effect size b_{xy}.. OR = $e^{b}{}_{xy}$. SNP inclusion threshold: $p \leq$ 1e-5.

Supplementary Figure 8. Betas for genome-wide significant SNPs from association analysis of educational attainment and intelligence test performance plotted against beta for voter turnout by sample



Note. The linkage disequilibrium structure between the SNPs was calculated from European individuals of the 1000 genomes phase 3 dataset. The X-axis denotes the effect sizes of the SNPs used in the GSMR analysis towards the exposure and Y-axis indicates the effect sizes of the same SNPs towards turnout. The blue trend line indicates the correlation between the effect sizes.

S9. Information and results from a bidirectional GSMR

As a further test of the relationship between PS for educational attainment and intelligence test performance, respectively, and voter turnout and to explore the direction of the relationship, we utilized the bidirectional summary statistics based Mendelian randomization analysis using the GSMR package of the GCTA software suite.

S9.1 Methods

The instrumental variables for the reverse GSMR analysis are the SNPs associated with electoral turnout, which are different from the SNPs used for the forward GSMR analysis where SNPs associated with educational attainment and intelligence test performance were used as instrumental variables. We reduce the genome wide significance threshold for selecting instrumental variables to $1x10^{-5}$ so that there are at least 15 independent SNPs to be used as instruments for each election and sample. For the European election, this threshold was further reduced to $1x10^{-4}$. The reverse GSMR analysis and the associated simulations for false positive rates has been described in detail (Zhu et al 2018).

Bidirectional GSMR was previously used to estimate the direction of effect between major depressive disorder and phenotypically associated traits (Wray et al 2018), height and coronary artery disease (Marouli et al 2019) and cannabis use and schizophrenia (Pasman et al 2018).

Linkage disequilibrium between SNPs was estimated from the European individuals of the 1000 genomes phase 3 dataset cohort genotypes with clumping $r^2 < 0.05$ and HEIDI outlier threshold at 0.01. SNPs with suspected pleiotropic effects were excluded using the HEIDI-outlier analysis implemented in the GSMR package.

S9.2. Results

The results from the bidirectional GSMR are reported in Supplementary Table 13 below and support our previous findings suggesting a relationship between genetic disposition towards educational attainment and performance on intelligence tests and turnout across both samples using PS for educational attainment and intelligence test performance respectively as exposure variables. These results suggest that the direction of the relationship is from genetic disposition towards educational attainment and performance on intelligence tests respectively to voter turnout.

Consistent with the suggested causal ordering between indicators of resources for politics and turnout in the resource model the analyses using turnout in any of the three elections as the exposure do not indicate that an increase in voter turnout has a causal effect on either educational attainment or performance on intelligence tests, with odds ratios varying between 0.99 (SE = 0.0045, p = 0.73) and 1.01 (SE = 0.0019, p = 0.35) in the nationally representative cohort and between 0.995 (SE = 0.0015, p = 0.0011) and 1.016 (SE = 0.0077, p = 0.08) in the psychiatric cases.

S9.3 Limitations

While the results of the bidirectional GSMR analyses support the findings in the main text and provide suggestive results regarding the direction of the relationships, the limitations to the MR methodology implies that we cannot determine whether these genetic correlations reflect causal effects of phenotypic education and phenotypic intelligence test performance on voter turnout.

We acknowledge that the bidirectional GSMR analyses reported in Supplementary Table 13 are statistically underpowered and the results in Supplementary Table 13 should therefore only be viewed as suggestive. Power calculations for GSMR, especially for binary outcomes are non-trivial. However, according to Zhu et. al (2018) in the paper introducing the method, it is mentioned that the power of GSMR to detect effect of one trait on another increases proportionally with an increase in the number of instrumental variables used in the analysis. While there have been examples in case-control samples where a bi-directional effect has been discovered in GSMR analysis with as few as 10 instrumental variables, most notably between BMI and type 2 diabetes, there are several other factors such as the number of strongly associated SNPs, sample size of the GWAS etc. that influence statistical power. We acknowledge that there is a disparity in the sample sizes of the summary statistics between our exposure and outcome traits, as well as the lack of genome wide significant SNPs when using voter turnout as exposure in the reverse GSMR, which limits our ability to reject the null hypothesis. A better powered GWAS in the future can hopefully address this limitation.

Supplementary Table 13. Results from the bidirectional Mendelian randomization analysis

Sample	Exposure	Turnout outcome	b_{xy}	OR	SE	p-value	nSNPs	HEIDI Outlier SNPs
Psychiatric	Educational attainment	National election 2015	0.867462	2.38	0.0557	1.18E-54	1557	5
	National election 2015	Educational attainment	0.008316	1.01	0.0036	0.0236	25	0
	IQ test performance	National election 2015	0.42646	1.53	0.0505	3.45E-17	771	1
	National election 2015	IQ test performance	-0.0034929	0.99	0.006	0.56	23	0
	Educational attainment	European election 2014	0.941797	2.56	0.0495	1.19E-80	2927	1
	European Election 2014	Educational attainment	-0.005114	0.99	0.0015	0.0011	134	2
	IQ test performance	European election 2014	0.53369	1.68	0.0407	3.88E-39	1771	2
	European election 2014	IQ test performance	0.0019	1	0.0026	0.46	123	2
	Educational attainment	Municipal election 2013	0.877919	2.41	0.0456	2.97E-82	1559	2
	Municipal election 2013	Educational attainment	0.00057	1	0.0047	0.9	19	0
	IQ test performance	Municipal election 2013	0.478133	1.61	0.0415	1.77E-30	770	2
	Municipal election 2013	IQ test performance	0.01357	1.01	0.0077	0.08	17	0
Nationally Repre-sentative	Educational attainment	National election 2015	1.08501	2.95	0.1001	2.46E-27	1562	2
	National election 2015	Educational attainment	-3.54E-05	0.99	0.0027	0.98	15	0
	IQ test performance	National election 2015	0.518502	1.67	0.0911	1.29E-08	770	3
	National election 2015	IQ test performance	-0.001562	0.99	0.0045	0.73	15	0
	Educational attainment	European election 2014	0.987598	2.68	0.0736	5.91E-41	2929	3
	European election 2014	Educational attainment	0.005464	1	0.0011	1.45E-06	116	2
	IQ test performance	European election 2014	0.492325	1.63	0.0606	4.93E-16	1770	3
	European election 2014	IQ test performance	0.001772	1.01	0.0019	0.35	108	2
	Educational attainment	Municipal election 2013	0.973022	2.64	0.0741	2.34E-39	1564	1
	Municipal election 2013	Educational attainment	0.005242	1	0.0033	0.11	17	0
	IQ test performance	Municipal election 2013	0.529408	1.69	0.0673	3.94E-15	773	0
	Municipal E election 2013	IQ test performance	0.007948	1	0.0057	0.16	15	0

Note. GWAS threshold: $p \leq$ 1e- 5 for national and municipal elections, p < 1e-4 for European election. b_{xy} refers to the effect size of the exposure on outcome. SE, p-value are respectively the standard error, associated p-value of the effect size b_{xy}.. OR = $e^{b_{xy}}$.

S10. References

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*, *62*(5), 1198-1211.

Bhatti, Y., Dahlgaard, J. O., Hansen, J. H., & Hansen, K. M. (2014a). Hvem stemte og hvem blev hjemme? Valgdeltagelsen ved kommunalvalget 19. november 2013. Beskrivende analyser af valgdeltagelsen baseret på registerdata. CVAP Working Papers Series CVAP 2/2014, Copenhagen: University of Copenhagen (2014a).

Bhatti, Y., Dahlgaard, J. O., Hansen, J. H., & Hansen, K. M. (2014b). Hvem stemte til EP-valget 2014? - Valgdeltagelsen ved Europa-Parlamentsvalget 25. maj 2014. Beskrivende analyser af valgdeltagelsen baseret på registerdata, CVAP Working Paper Series 4/2014, Copenhagen: University of Copenhagen.

Bhatti, Y., Dahlgaard, J. O., Hansen, J. H., & Hansen, K. M. (2016). Valgdeltagelsen og vælgerne til Folketingsvalget 2015, CVAP Working Paper Series 1/2016, Copenhagen: University of Copenhagen (2016).

Canty A, Ripley BD (2020). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-25.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge. ISBN 0-521-57391-2, http://statwww.epfl.ch/davison/BMA/.

De Moor, M. H., Costa, P. T., Terracciano, A., Krueger, R. F., De Geus, E. J., Toshiko, T., ... & Amin, N. (2012). Meta-analysis of genome-wide association studies for personality. *Molecular psychiatry*, *17*(3), 337-349.

Duan, S., Zhang, W., Cox, N. J., & Dolan, M. E. (2008). FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation*, *3*(3), 139.

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* Pearson Prentice Hall.

Lee, S. H., Goddard, M. E., Wray, N. R., & Visscher, P. M. A. Better coefficient of determination for genetic profile analysis. *Genetic epidemiology*, *36*(3), 214-224 (2012).

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

Marouli, E., Del Greco, M. F., Astley, C. M., Yang, J., Ahmad, S., Berndt, S. I., ... & van Vliet-Ostaptchouk, J. V. (2019). Mendelian randomisation analyses find pulmonary factors mediate the effect of height on coronary artery disease. *Communications Biology*, *2*(1), 119.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Pasman, J. A., Verweij, K. J., Gerring, Z., Stringer, S., Sanchez-Roige, S., Treur, J. L., ... & Ip, H. F. (2018). GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal effect of schizophrenia liability. *Nature neuroscience*, *21*(9), 1161-1170.

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*(262-273), 1-47.

Pedersen, C. B. (2011). The Danish civil registration system. *Scandinavian journal of public health*, *39*(7_suppl), 22-25.

Pedersen, C. B., Bybjerg-Grauholm, J., Pedersen, M. G., Grove, J., Agerbo, E., Baekvad-Hansen, M., ... & Goldstein, J. I. (2018). The iPSYCH2012 case–cohort sample: new directions for

unravelling genetic and environmental architectures of severe mental disorders. *Molecular psychiatry*, *23*(1), 6-14.
Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., ... & Grasby, K. L. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics*, *50*(7), 912.
Schork, A. J., Won, H., Appadurai, V., Nudel, R., Gandal, M., Delaneau, O., ... & Pedersen, M. G. (2019). A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nature neuroscience*, *22*(3), 353.
Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., ... & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, *48*(5), 481.
Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... & Buchkovich, M. L. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, *46*(11), 1173.
Wray, N. R., & Gottesman, I. I. (2012). Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Frontiers in genetics*, *3*, 118: 1-12.
Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., ... & Bacanu, S. A. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, *50*(5), 668.

Election	Sample	trait2	PMID	Category	ethnicity	note	rg	se	z	p	h2_obs	h2_obs_se	h2_int	h2_int_se	gcov_int	gcov_int_se
Municipal 2013	Psychiatric Sample	Age of smoking initiation	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1082	0.1473	-0.7346	0.4626	0.0645	0.0185	0.9955	0.0072	0.0133	0.0053
Municipal 2013	Psychiatric Sample	Child birth length	25281659	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1451	0.109	1.3308	0.1833	0.1664	0.0271	0.9914	0.0091	-0.003	0.0063
Municipal 2013	Psychiatric Sample	Child birth weight	23202124	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1094	0.1314	0.8322	0.4053	0.108	0.0185	1.003	0.0068	0.0002	0.0056
Municipal 2013	Psychiatric Sample	Body mass index	20935630	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0734	0.0531	-1.3817	0.1671	0.1889	0.0096	0.9995	0.0125	-0.0003	0.0055
Municipal 2013	Psychiatric Sample	Body fat	26833246	anthropometric	Mixed	Caution: using this data may yield less robust results due to minor departure of th	-0.2713	0.0696	-3.8965	9.76E-05	0.1044	0.0083	0.9037	0.0073	0.0112	0.005
Municipal 2013	Psychiatric Sample	Childhood IQ	23358156	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5869	0.1284	4.5703	4.87E-06	0.2836	0.0473	0.9985	0.0104	-0.0061	0.0064
Municipal 2013	Psychiatric Sample	Childhood obesity	22484627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0598	0.0832	-0.7184	0.4725	0.4362	0.0418	0.9181	0.0078	-0.0019	0.0055
Municipal 2013	Psychiatric Sample	Cigarettes smoked per day	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3564	0.1359	-2.6223	0.0087	0.0524	0.0155	1.0099	0.0077	0.0004	0.005
Municipal 2013	Psychiatric Sample	Depressive symptoms	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.288	0.0772	-3.729	0.0002	0.0473	0.0042	1.0026	0.0094	0.0017	0.0059
Municipal 2013	Psychiatric Sample	Years of schooling 2016	27225129	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6414	0.0626	10.2479	1.21E-24	0.1261	0.0049	0.9246	0.0123	-0.0022	0.0063
Municipal 2013	Psychiatric Sample	Extreme bmi	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0743	0.093	-0.7994	0.4241	0.6876	0.0605	1.0281	0.0138	0.0039	0.0088
Municipal 2013	Psychiatric Sample	Former vs Current smoker	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3357	0.1398	2.4007	0.0164	0.0596	0.0116	1.0036	0.0078	-0.0024	0.0056
Municipal 2013	Psychiatric Sample	Anorexia Nervosa	24514567	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0416	0.0775	-0.5365	0.5916	0.4493	0.0332	0.9379	0.0086	0.0118	0.0058
Municipal 2013	Psychiatric Sample	Extreme height	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.081	0.0729	1.1108	0.2667	1.2949	0.1154	1.0013	0.0211	0.0042	0.0091
Municipal 2013	Psychiatric Sample	Extreme waist-to-hip ratio	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0925	0.1245	-0.7428	0.4576	0.3433	0.062	0.98	0.0095	0.0023	0.0076
Municipal 2013	Psychiatric Sample	Height_2010	20881960	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1375	0.0511	2.693	0.0071	0.284	0.0161	0.9959	0.0185	0.002	0.0073
Municipal 2013	Psychiatric Sample	Obesity class 1	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0577	0.055	-1.0491	0.2942	0.2139	0.0117	1.013	0.0133	-0.0085	0.007
Municipal 2013	Psychiatric Sample	Obesity class 2	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0459	0.0756	-0.608	0.5432	0.1747	0.0134	1.0094	0.0128	-0.0037	0.0077
Municipal 2013	Psychiatric Sample	Obesity class 3	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0728	0.1026	-0.7095	0.478	0.1153	0.0155	0.9826	0.0111	0.0066	0.0075
Municipal 2013	Psychiatric Sample	Overweight	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1047	0.0585	-1.79	0.0735	0.1086	0.0066	1.0188	0.0109	-0.0015	0.0059
Municipal 2013	Psychiatric Sample	Neo-conscientiousness	21173776	personality	European	Caution: using this data may yield results outside bounds due to relative low Z sc	0.1378	0.1661	0.8297	0.4067	0.0799	0.033	0.9987	0.0086	-0.004	0.0062
Municipal 2013	Psychiatric Sample	Neo-openness to experience	21173776	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3686	0.1464	2.5177	0.0118	0.1181	0.0284	0.9885	0.0077	-0.0096	0.006
Municipal 2013	Psychiatric Sample	Hip circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0214	0.0507	-0.4225	0.6727	0.1287	0.0059	0.8525	0.0095	0.0047	0.0052
Municipal 2013	Psychiatric Sample	Alzheimers disease	24162737	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1062	0.145	0.7325	0.4639	0.0441	0.0195	1.0656	0.0281	-0.0074	0.0062
Municipal 2013	Psychiatric Sample	Infant head circumference	22504419	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2842	0.1291	2.2017	0.0277	0.2392	0.0449	0.9882	0.0078	-0.0063	0.0062
Municipal 2013	Psychiatric Sample	Age at Menarche	25231870	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0633	0.0521	-1.2145	0.2246	0.2085	0.0104	0.9288	0.0135	0.0002	0.0066
Municipal 2013	Psychiatric Sample	Age at Menopause	26414677	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1184	0.0706	1.6764	0.0937	0.1368	0.0166	0.9856	0.0135	-0.0019	0.0055
Municipal 2013	Psychiatric Sample	Neuroticism	27089181	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1765	0.0648	-2.7264	0.0064	0.0912	0.0079	0.9791	0.0149	0.0006	0.0063
Municipal 2013	Psychiatric Sample	Attention deficit hyperactivity diso	20732625	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z sc	-0.58	0.1982	-2.926	0.0034	0.2354	0.0974	1.0071	0.0077	0.0038	0.0049
Municipal 2013	Psychiatric Sample	Bipolar disorder	21926972	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.033	0.0806	0.4087	0.6828	0.4254	0.0383	1.0232	0.0083	0.0029	0.0055
Municipal 2013	Psychiatric Sample	PGC cross-disorder analysis	23453885	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1506	0.082	-1.8354	0.0664	0.1749	0.0132	1.0095	0.0121	0.0038	0.0067
Municipal 2013	Psychiatric Sample	Major depressive disorder	22472876	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0009	0.1153	-0.0081	0.9936	0.1792	0.027	1.0001	0.0068	-0.0026	0.0054
Municipal 2013	Psychiatric Sample	Autism spectrum disorder	0	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.139	0.0902	1.5413	0.1232	0.4972	0.0566	0.9569	0.0078	-0.0064	0.0053
Municipal 2013	Psychiatric Sample	Parkinsons disease	19915575	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1067	0.1001	-1.0663	0.2863	0.3178	0.1354	1.13	0.0114	0.0041	0.0051
Municipal 2013	Psychiatric Sample	Schizophrenia	25056061	psychiatric	Mixed	Caution: using this data may yield less robust results due to minor departure of th	-0.169	0.0577	-2.9277	0.0034	0.4614	0.0196	1.0464	0.0148	-0.0125	0.0076
Municipal 2013	Psychiatric Sample	College completion	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6351	0.0895	7.0985	1.26E-12	0.0788	0.0062	1.0202	0.0098	0.0027	0.0063
Municipal 2013	Psychiatric Sample	Subjective well being	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.015	0.0844	-0.1778	0.8589	0.0244	0.0023	1.003	0.0089	0.0125	0.0066
Municipal 2013	Psychiatric Sample	Ever vs never smoked	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3623	0.0983	-3.6853	0.0002	0.07	0.0069	1.0028	0.0072	0.0014	0.0053
Municipal 2013	Psychiatric Sample	Waist circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1375	0.049	-2.8085	0.005	0.1222	0.0053	0.8378	0.0095	0.0097	0.0049
Municipal 2013	Psychiatric Sample	Waist-to-hip ratio	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2073	0.0586	-3.5386	0.0004	0.1116	0.0073	0.9147	0.0106	0.0072	0.0055
Municipal 2013	Psychiatric Sample	Amyotrophic lateral sclerosis	27455348	neurological	European	Caution: using this data may yield results outside bounds due to relative low Z sc	-0.205	0.151	-1.3572	0.1747	0.057	0.0135	0.9873	0.0069	0.0067	0.0053
Municipal 2013	Psychiatric Sample	Difference in height between adol	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1796	0.128	-1.4033	0.1605	0.4896	0.105	0.9755	0.0071	0.0056	0.0053
Municipal 2013	Psychiatric Sample	Parents age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2113	0.1467	1.4401	0.1498	0.0288	0.0076	1.015	0.0073	-0.0008	0.0058
Municipal 2013	Psychiatric Sample	Mothers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4971	0.1341	3.7075	0.0002	0.0372	0.0079	1.0094	0.008	-0.0063	0.0057
Municipal 2013	Psychiatric Sample	Fathers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1484	0.116	1.2787	0.201	0.04	0.007	1.0161	0.0078	0.0047	0.0058
Municipal 2013	Psychiatric Sample	Difference in height between child	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0989	0.1166	-0.8486	0.3961	0.3496	0.0567	0.9662	0.0091	0.0086	0.0061
Municipal 2013	Psychiatric Sample	Sitting height ratio	25865494	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1045	0.0929	1.1251	0.2606	0.2263	0.0293	0.9789	0.0089	-0.0082	0.0057
Municipal 2013	Psychiatric Sample	Height; Females at age 10 and ma	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2239	0.0954	2.3478	0.0189	0.4112	0.0477	0.9555	0.0092	0.0001	0.0058
Municipal 2013	Psychiatric Sample	Sleep duration	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1258	0.0886	1.4194	0.1558	0.0558	0.0051	1.0193	0.0085	-0.0007	0.0059
Municipal 2013	Psychiatric Sample	Birth weight	27680694	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1294	0.0617	2.0991	0.0358	0.1031	0.0069	1.0275	0.0096	-0.0024	0.0053
Municipal 2013	Psychiatric Sample	Chronotype	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0775	0.0673	1.1513	0.2496	0.1051	0.0065	1.0043	0.0093	-0.0133	0.006
Municipal 2013	Psychiatric Sample	Years of schooling (proxy cognitive	25201988	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6569	0.0794	8.2762	1.27E-16	0.1082	0.0081	1.0216	0.0108	0.0018	0.0066
Municipal 2013	Psychiatric Sample	Years of schooling 2013	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6268	0.076	8.246	1.64E-16	0.0865	0.0067	1.0104	0.0107	0.0016	0.0061
Municipal 2013	Psychiatric Sample	Neuroticism	24828478	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1793	0.1301	-1.3786	0.168	0.0139	0.0034	1.0116	0.0073	-0.0052	0.0052
Municipal 2013	Psychiatric Sample	Age of first birth	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5757	0.0717	8.0284	9.87E-16	0.0619	0.0038	0.9529	0.0097	0.0008	0.0058
Municipal 2013	Psychiatric Sample	Number of children ever born	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2494	0.0752	-3.3177	0.0009	0.0254	0.0019	0.9699	0.008	0.0039	0.0055
Municipal 2013	Psychiatric Sample	Attention deficit hyperactivity diso	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z sc	-0.1993	0.2222	-0.8969	0.3698	0.0719	0.0295	0.9921	0.0081	-0.0042	0.0066
Municipal 2013	Psychiatric Sample	Attention deficit hyperactivity diso	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z sc	-0.1985	0.2216	-0.8959	0.3703	0.0732	0.0299	1.0066	0.0082	-0.0043	0.0067
Municipal 2013	Psychiatric Sample	Intelligence	28530673	cognitive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4976	0.0781	6.3724	1.86E-10	0.1985	0.0112	0.9926	0.0098	-0.0037	0.0063
Municipal 2013	Psychiatric Sample	Insomnia	28604731	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.329	0.0956	-3.4413	0.0006	0.0484	0.0052	0.9997	0.0079	0.0027	0.0054
Municipal 2013	Psychiatric Sample	Insomnia	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2982	0.0841	-3.5465	0.0004	0.1373	0.0129	0.9972	0.0091	0.0045	0.0056
Municipal 2013	Psychiatric Sample	Excessive daytime sleepiness	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0286	0.098	0.2922	0.7701	0.051	0.0052	1.0112	0.0082	0.0005	0.0059
Municipal 2013	Psychiatric Sample	Cigarettes Per Day	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z sc	-0.0294	0.1544	-0.1901	0.8492	0.0155	0.0082	1.0017	0.0144	-0.009	0.0073
Municipal 2013	Psychiatric Sample	Smoking Cessation	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z sc	0.6386	0.6883	0.9278	0.3535	0.0088	0.0033	1.0162	0.0233	-0.0073	0.0186
Municipal 2013	Psychiatric Sample	Smoking Initiation	30617275	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3207	0.1695	-1.8922	0.0585	0.0036	0.0011	1.09	0.0106	-0.0044	0.0064
Municipal 2013	Psychiatric Sample	Pack Years	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z sc	NA	NA	NA	NA	0.1119	0.0448	0.9704	0.0893	-0.0776	0.0707
Municipal 2013	Psychiatric Sample	Own birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1389	0.0546	2.5445	0.0109	0.0978	0.0054	1.0592	0.0149	-0.0099	0.0065
Municipal 2013	Psychiatric Sample	Own birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1406	0.0529	2.6561	0.0079	0.0938	0.005	1.0478	0.0144	-0.0116	0.0064
Municipal 2013	Psychiatric Sample	Own birth weight (fetal effect) adj	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0306	0.0831	0.3676	0.7132	0.0298	0.0029	1.0279	0.0096	0.0011	0.0059
Municipal 2013	Psychiatric Sample	Offspring birth weight (maternal e	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2048	0.0597	3.4322	0.0006	0.1036	0.0061	1.0395	0.0118	-0.0184	0.0066
Municipal 2013	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2023	0.0587	3.4485	0.0006	0.0997	0.0055	1.0234	0.0118	-0.0183	0.0065
Municipal 2013	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2327	0.085	2.737	0.0062	0.0293	0.0026	1.0191	0.0089	-0.0127	0.0062

Election	Cohort	trait2	PMID	Category	ethnicity	note	rg	se	z	p	h2_obs	h2_obs_se	h2_int	h2_int_se	gcov_int	gcov_int_se
Municipal 2013	National Representative	Age of smoking initiation	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5265	0.4518	1.1652	0.2439	0.0634	0.0186	0.996	0.0073	-1.62E-05	0.0051
Municipal 2013	National Representative	Child birth length	25281659	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.664	0.5148	1.2898	0.1971	0.1663	0.027	0.9914	0.0091	-0.0027	0.0061
Municipal 2013	National Representative	Child birth weight	23202124	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3149	0.2672	1.1785	0.2386	0.1068	0.0185	1.0037	0.0068	2.96E-05	0.0054
Municipal 2013	National Representative	Body mass index	20935630	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1468	0.1384	-1.0603	0.289	0.1887	0.0096	0.9998	0.0125	0.0002	0.0067
Municipal 2013	National Representative	Body fat	26833246	anthropometric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; S	-0.4526	0.2103	-2.1519	0.0314	0.1044	0.0083	0.9037	0.0073	-0.0011	0.0054
Municipal 2013	National Representative	Childhood IQ	23358156	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6927	0.4448	1.5574	0.1194	0.283	0.0473	0.9986	0.0104	0.0058	0.0065
Municipal 2013	National Representative	Childhood obesity	22484627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3845	0.2301	-1.6713	0.0947	0.4358	0.0416	0.9181	0.0077	0.0065	0.0051
Municipal 2013	National Representative	Cigarettes smoked per day	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4219	0.3573	-1.1807	0.2377	0.0524	0.0154	1.0099	0.0077	-0.0044	0.0053
Municipal 2013	National Representative	Depressive symptoms	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.6192	0.4298	-1.4409	0.1496	0.0473	0.0042	1.0026	0.0094	-0.0043	0.0059
Municipal 2013	National Representative	Years of schooling 2016	27225129	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.951	0.3543	2.6842	0.0073	0.1261	0.0049	0.9243	0.0122	-0.0064	0.0055
Municipal 2013	National Representative	Extreme bmi	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.014	0.344	-0.0406	0.9676	0.6876	0.0605	1.0281	0.0138	-0.0036	0.0089
Municipal 2013	National Representative	Former vs Current smoker	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.8719	0.5007	1.7414	0.0816	0.0597	0.0116	1.0035	0.0078	-0.0015	0.0053
Municipal 2013	National Representative	Anorexia Nervosa	24514567	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0514	0.1888	-0.2725	0.7852	0.4491	0.0332	0.938	0.0086	0.001	0.0057
Municipal 2013	National Representative	Extreme height	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5114	0.5789	0.8834	0.377	1.2949	0.1154	1.0013	0.0211	-0.0095	0.0096
Municipal 2013	National Representative	Extreme waist-to-hip ratio	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1586	0.4294	-0.3694	0.7119	0.3433	0.062	0.98	0.0095	-0.0043	0.0077
Municipal 2013	National Representative	Height_2010	20881960	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2887	0.1363	2.119	0.0341	0.2837	0.0161	0.9964	0.0185	-0.0008	0.0077
Municipal 2013	National Representative	Obesity class 1	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4336	0.4015	-1.08	0.2802	0.2139	0.0117	1.013	0.0133	0.0051	0.0075
Municipal 2013	National Representative	Obesity class 2	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.5756	0.6243	-0.922	0.3565	0.1747	0.0134	1.0094	0.0128	0.0123	0.0083
Municipal 2013	National Representative	Obesity class 3	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.7831	0.4907	-1.5958	0.1105	0.1153	0.0155	0.9826	0.0111	0.0097	0.0075
Municipal 2013	National Representative	Overweight	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1636	0.159	-1.0294	0.3033	0.1086	0.0066	1.0189	0.0109	-0.0038	0.0062
Municipal 2013	National Representative	Neo-conscientiousness	21173776	personality	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.0192	0.4426	-0.0434	0.9654	0.0799	0.033	0.9987	0.0086	0.0061	0.0054
Municipal 2013	National Representative	Neo-openness to experience	21173776	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0268	0.3726	0.0718	0.9427	0.1182	0.0284	0.9885	0.0077	0.0086	0.006
Municipal 2013	National Representative	Hip circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0801	0.1418	-0.565	0.572	0.1289	0.0059	0.8517	0.0094	0.003	0.0055
Municipal 2013	National Representative	Alzheimers disease	24162737	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4438	0.3394	-1.3075	0.191	0.0439	0.0195	1.0658	0.0281	0.0009	0.0053
Municipal 2013	National Representative	Infant head circumference	22504419	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4454	0.3582	1.2436	0.2137	0.2393	0.0449	0.9882	0.0077	-0.0002	0.0057
Municipal 2013	National Representative	Age at Menarche	25231870	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1428	0.1287	1.1094	0.2673	0.2083	0.0104	0.9293	0.0134	-0.0077	0.0064
Municipal 2013	National Representative	Age at Menopause	26414677	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1318	0.1525	-0.864	0.3876	0.1363	0.0166	0.9862	0.0135	0.0069	0.0054
Municipal 2013	National Representative	Neuroticism	27089181	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0934	0.2078	-0.4495	0.6531	0.0912	0.0079	0.9791	0.0149	-0.016	0.0064
Municipal 2013	National Representative	Attention deficit hyperactivity disorder	20732625	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.2841	0.3914	-0.726	0.4678	0.2324	0.0972	1.0076	0.0076	0.0024	0.0054
Municipal 2013	National Representative	Bipolar disorder	21926972	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2044	0.2157	0.9478	0.3432	0.4255	0.0381	1.0233	0.0083	0.0005	0.0058
Municipal 2013	National Representative	PGC cross-disorder analysis	23453885	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1396	0.1733	0.8054	0.4206	0.1739	0.0132	1.0108	0.0121	-0.0051	0.0074
Municipal 2013	National Representative	Major depressive disorder	22472876	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2967	0.2361	-1.2568	0.2088	0.1802	0.0272	0.9998	0.0068	0.0035	0.0051
Municipal 2013	National Representative	Autism spectrum disorder	0	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2696	0.2029	1.3289	0.1839	0.4982	0.0567	0.9568	0.0078	0.0011	0.0046
Municipal 2013	National Representative	Parkinsons disease	19915575	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4206	0.367	1.1462	0.2517	0.3178	0.1354	1.13	0.0114	-0.0086	0.005
Municipal 2013	National Representative	Schizophrenia	25056061	psychiatric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; S	-0.2285	0.1341	-1.7043	0.0883	0.4616	0.0196	1.0462	0.0148	-0.0045	0.0067
Municipal 2013	National Representative	College completion	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.8889	0.3708	2.3972	0.0165	0.0788	0.0062	1.0202	0.0098	0.0033	0.006
Municipal 2013	National Representative	Subjective well being	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1535	0.272	0.5645	0.5724	0.0244	0.0023	1.003	0.0089	0.0014	0.0059
Municipal 2013	National Representative	Ever vs never smoked	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.5659	0.3504	-1.615	0.1063	0.0703	0.0069	1.0023	0.0072	-0.0074	0.0058
Municipal 2013	National Representative	Waist circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3171	0.1912	-1.6589	0.0971	0.1223	0.0053	0.8374	0.0094	0.0067	0.006
Municipal 2013	National Representative	Waist-to-hip ratio	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3943	0.1843	-2.1397	0.0324	0.1114	0.0073	0.9151	0.0106	0.0048	0.0058
Municipal 2013	National Representative	Amyotrophic lateral sclerosis	27455348	neurological	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.4507	0.3654	-1.2333	0.2175	0.0571	0.0135	0.9872	0.007	0.0056	0.0054
Municipal 2013	National Representative	Difference in height between adolescence and	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0118	0.2642	-0.0446	0.9644	0.4872	0.1048	0.9757	0.007	0.0003	0.0049
Municipal 2013	National Representative	Parents age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5263	0.3369	1.5622	0.1182	0.0287	0.0075	1.0153	0.0073	0.0002	0.0054
Municipal 2013	National Representative	Mothers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4539	0.3327	1.3642	0.1725	0.0377	0.008	1.0087	0.008	-0.0026	0.0053
Municipal 2013	National Representative	Fathers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7183	0.3481	2.0638	0.039	0.0403	0.007	1.0156	0.0078	-0.0061	0.0053
Municipal 2013	National Representative	Difference in height between childhood and a	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1264	0.2543	0.4972	0.6191	0.3488	0.0567	0.9664	0.0091	0.0003	0.0057
Municipal 2013	National Representative	Sitting height ratio	25865494	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2139	0.2329	0.9185	0.3584	0.2277	0.0292	0.9782	0.0088	-0.0051	0.0055
Municipal 2013	National Representative	Height; Females at age 10 and males at age 1	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0441	0.1923	0.2296	0.8184	0.4109	0.0477	0.9556	0.0092	0.0034	0.0056
Municipal 2013	National Representative	Sleep duration	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1448	0.1989	0.7279	0.4667	0.056	0.0051	1.0185	0.0086	-0.0026	0.0056
Municipal 2013	National Representative	Birth weight	27680694	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3118	0.1857	1.6796	0.093	0.103	0.0069	1.0275	0.0096	-0.0075	0.0059
Municipal 2013	National Representative	Chronotype	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1478	0.1671	0.8842	0.3766	0.105	0.0065	1.0045	0.0092	-0.0041	0.0059
Municipal 2013	National Representative	Years of schooling (proxy cognitive performan	25201988	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.8985	0.3917	2.2936	0.0218	0.1082	0.0081	1.0216	0.0108	0.001	0.0068
Municipal 2013	National Representative	Years of schooling 2013	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7815	0.3413	2.2898	0.022	0.0865	0.0067	1.0104	0.0107	0.004	0.0064
Municipal 2013	National Representative	Neuroticism	24828478	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3819	0.4054	-0.9422	0.3461	0.0138	0.0034	1.0116	0.0073	-0.0116	0.0053
Municipal 2013	National Representative	Age of first birth	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.9186	0.3759	2.444	0.0145	0.0617	0.0038	0.9539	0.0097	-0.0031	0.0058
Municipal 2013	National Representative	Number of children ever born	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2788	0.1976	-1.4111	0.1582	0.0254	0.0019	0.97	0.008	0.0038	0.0051
Municipal 2013	National Representative	Attention deficit hyperactivity disorder (GC)	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-1.4905	1.0352	-1.4398	0.1499	0.0719	0.0295	0.9921	0.0081	0.0083	0.0059
Municipal 2013	National Representative	Attention deficit hyperactivity disorder (No GC	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-1.4736	1.0127	-1.4551	0.1456	0.0732	0.0299	1.0066	0.0082	0.0084	0.0059
Municipal 2013	National Representative	Intelligence	28530673	cognitive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7823	0.2918	2.6812	0.0073	0.1983	0.0113	0.9928	0.0098	-0.0025	0.0055
Municipal 2013	National Representative	Insomnia	28604731	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.6322	0.2986	-2.1173	0.0342	0.0484	0.0052	0.9997	0.0079	0.0052	0.0057
Municipal 2013	National Representative	Insomnia	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4977	0.2505	-1.9868	0.0469	0.1374	0.0128	0.9971	0.009	0.0064	0.0058
Municipal 2013	National Representative	Excessive daytime sleepiness	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.016	0.2062	-0.0773	0.9383	0.051	0.0052	1.0112	0.0082	5.51E-05	0.0051
Municipal 2013	National Representative	Cigarettes Per Day	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.7618	0.3699	-2.0595	0.0394	0.0157	0.0081	1.0009	0.0143	0.0143	0.0063
Municipal 2013	National Representative	Smoking Cessation	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.5271	0.4912	1.0731	0.2832	0.0088	0.0033	1.0164	0.023	0.0038	0.0165
Municipal 2013	National Representative	Smoking Initiation	30617275	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1949	0.2693	-0.7239	0.4691	0.0036	0.0011	1.0903	0.0107	-0.0027	0.0066
Municipal 2013	National Representative	Pack Years	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Municipal 2013	National Representative	Own birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.29	0.1647	1.7613	0.0782	0.0978	0.0054	1.0593	0.0149	-0.0092	0.0069
Municipal 2013	National Representative	Own birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2754	0.164	1.6787	0.0932	0.0938	0.0051	1.0475	0.0144	-0.0078	0.0068
Municipal 2013	National Representative	Own birth weight (fetal effect) adjusted for m	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.157	0.2026	0.7747	0.4385	0.0297	0.0029	1.0283	0.0097	-0.0025	0.006
Municipal 2013	National Representative	Offspring birth weight (maternal effect) adjus	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3215	0.1595	2.0155	0.0439	0.1034	0.0061	1.0401	0.0117	-0.0106	0.006
Municipal 2013	National Representative	Offspring birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3059	0.1597	1.9152	0.0555	0.0995	0.0055	1.0241	0.0117	-0.0094	0.0061
Municipal 2013	National Representative	Offspring birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3416	0.1894	1.8039	0.0712	0.0292	0.0026	1.0197	0.0089	-0.007	0.0055

Election	Cohort	trait2	PMID	Category	ethnicity	note	rg	se	z	p	h2_obs	h2_obs_se	h2_int	h2_int_se	gcov_int	gcov_int_se
European 2014	Psychiatric Sample	Age of smoking initiation	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1615	0.177	0.9128	0.3613	0.0642	0.0185	0.9956	0.0072	0.0054	0.0053
European 2014	Psychiatric Sample	Child birth length	25281659	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0137	0.1055	0.1303	0.8963	0.1664	0.0271	0.9914	0.0091	0.0063	0.006
European 2014	Psychiatric Sample	Child birth weight	23202124	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0249	0.1277	-0.1951	0.8453	0.1085	0.0185	1.0027	0.0068	0.0067	0.0057
European 2014	Psychiatric Sample	Body mass index	20935630	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0318	0.0602	-0.5285	0.5972	0.1889	0.0096	0.9993	0.0125	-0.0062	0.0063
European 2014	Psychiatric Sample	Body fat	26833246	anthropometric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; SI	-0.1867	0.0856	-2.18	0.0293	0.1045	0.0083	0.9037	0.0073	0.0027	0.0055
European 2014	Psychiatric Sample	Childhood IQ	23358156	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5294	0.1361	3.8889	0.0001	0.283	0.0473	0.9986	0.0104	-0.0002	0.0066
European 2014	Psychiatric Sample	Childhood obesity	22484627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.108	0.0803	-1.3458	0.1784	0.4357	0.0419	0.9182	0.0077	-0.0012	0.0049
European 2014	Psychiatric Sample	Cigarettes smoked per day	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.5066	0.1623	-3.1212	0.0018	0.0518	0.0155	1.0104	0.0077	0.0063	0.0049
European 2014	Psychiatric Sample	Depressive symptoms	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2153	0.0787	-2.7374	0.0062	0.0473	0.0042	1.0026	0.0094	-0.0028	0.0055
European 2014	Psychiatric Sample	Years of schooling 2016	27225129	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.624	0.064	9.755	1.76E-22	0.1261	0.0049	0.9246	0.0122	-0.0025	0.0057
European 2014	Psychiatric Sample	Extreme bmi	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0385	0.091	-0.423	0.6723	0.6876	0.0605	1.0281	0.0138	-0.0005	0.009
European 2014	Psychiatric Sample	Former vs Current smoker	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.406	0.1422	2.8553	0.0043	0.0596	0.0116	1.0037	0.0078	-0.0028	0.0048
European 2014	Psychiatric Sample	Anorexia Nervosa	24514567	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0632	0.0778	-0.8122	0.4167	0.4492	0.0332	0.9379	0.0086	0.0091	0.0058
European 2014	Psychiatric Sample	Extreme height	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0934	0.0728	1.2835	0.1993	1.2949	0.1154	1.0013	0.0211	0.0019	0.0084
European 2014	Psychiatric Sample	Extreme waist-to-hip ratio	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0322	0.1405	-0.2289	0.819	0.3433	0.062	0.98	0.0095	-0.006	0.0077
European 2014	Psychiatric Sample	Height_2010	20881960	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1125	0.0527	2.1372	0.0326	0.2838	0.0161	0.9963	0.0186	0.0002	0.0068
European 2014	Psychiatric Sample	Obesity class 1	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0125	0.0618	-0.2024	0.8396	0.2139	0.0117	1.013	0.0133	-0.0148	0.007
European 2014	Psychiatric Sample	Obesity class 2	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0004	0.0706	-0.0058	0.9954	0.1747	0.0134	1.0094	0.0128	-0.0141	0.0071
European 2014	Psychiatric Sample	Obesity class 3	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0225	0.1118	-0.2016	0.8402	0.1153	0.0155	0.9826	0.0111	0.0016	0.0076
European 2014	Psychiatric Sample	Overweight	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	2.99E-08	0.0633	4.72E-07	1	0.1086	0.0066	1.0188	0.0109	-0.0115	0.0063
European 2014	Psychiatric Sample	Neo-conscientiousness	21173776	personality	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.2806	0.1973	1.4224	0.1549	0.08	0.033	0.9986	0.0086	-0.0045	0.0057
European 2014	Psychiatric Sample	Neo-openness to experience	21173776	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6405	0.1676	3.8204	0.0001	0.1182	0.0284	0.9884	0.0077	-0.0096	0.0058
European 2014	Psychiatric Sample	Hip circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.001	0.0622	-0.0155	0.9876	0.1285	0.0059	0.8529	0.0095	0.0031	0.0058
European 2014	Psychiatric Sample	Alzheimers disease	24162737	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1599	0.1496	-1.0687	0.2852	0.0437	0.0195	1.0661	0.0281	0.0006	0.006
European 2014	Psychiatric Sample	Infant head circumference	22504419	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0696	0.1461	0.4766	0.6336	0.2391	0.0449	0.9883	0.0078	0.0032	0.006
European 2014	Psychiatric Sample	Age at Menarche	25231870	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0153	0.0484	-0.3148	0.7529	0.2084	0.0104	0.929	0.0134	-0.0015	0.0061
European 2014	Psychiatric Sample	Age at Menopause	26414677	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1581	0.0868	1.8215	0.0685	0.1365	0.0166	0.986	0.0135	-0.0041	0.0065
European 2014	Psychiatric Sample	Neuroticism	27089181	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1808	0.068	-2.6594	0.0078	0.0912	0.0079	0.9791	0.0149	0.0002	0.0069
European 2014	Psychiatric Sample	Attention deficit hyperactivity disorder	20732625	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5175	0.1945	-2.661	0.0078	0.2336	0.0974	1.0074	0.0077	0.0047	0.0049
European 2014	Psychiatric Sample	Bipolar disorder	21926972	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1006	0.0919	1.0946	0.2737	0.4248	0.0383	1.0234	0.0083	0.0009	0.0058
European 2014	Psychiatric Sample	PGC cross-disorder analysis	23453885	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0138	0.0868	0.1585	0.874	0.1754	0.0132	1.0089	0.012	-0.009	0.0079
European 2014	Psychiatric Sample	Major depressive disorder	22472876	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.041	0.1237	-0.3316	0.7402	0.1794	0.027	1.0001	0.0068	-0.0063	0.0056
European 2014	Psychiatric Sample	Autism spectrum disorder	0	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1107	0.1022	1.0826	0.279	0.4976	0.0565	0.9568	0.0077	0.0014	0.0056
European 2014	Psychiatric Sample	Parkinsons disease	19915575	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1101	0.1293	-0.852	0.3942	0.3178	0.1354	1.13	0.0114	0.0049	0.0058
European 2014	Psychiatric Sample	Schizophrenia	25056061	psychiatric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; SI	-0.0929	0.0601	-1.5455	0.1222	0.4616	0.0196	1.0462	0.0148	-0.0028	0.0082
European 2014	Psychiatric Sample	College completion	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6474	0.0914	7.0812	1.43E-12	0.0788	0.0062	1.0202	0.0098	-0.004	0.0063
European 2014	Psychiatric Sample	Subjective well being	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2194	0.0893	2.4569	0.014	0.0244	0.0023	1.003	0.0089	-0.0128	0.0063
European 2014	Psychiatric Sample	Ever vs never smoked	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3647	0.1149	-3.1744	0.0015	0.0701	0.0069	1.0027	0.0072	9.75E-05	0.0058
European 2014	Psychiatric Sample	Waist circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.107	0.0612	-1.7482	0.0804	0.1221	0.0052	0.838	0.0094	0.0016	0.0053
European 2014	Psychiatric Sample	Waist-to-hip ratio	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1775	0.0708	-2.5052	0.0122	0.1114	0.0074	0.9152	0.0107	-0.0014	0.0061
European 2014	Psychiatric Sample	Amyotrophic lateral sclerosis	27455348	neurological	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.2845	0.1602	-1.7759	0.0757	0.0569	0.0135	0.9874	0.007	0.004	0.0058
European 2014	Psychiatric Sample	Difference in height between adolescence an	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0037	0.1261	0.0291	0.9768	0.4869	0.105	0.9757	0.0071	0.0029	0.0049
European 2014	Psychiatric Sample	Parents age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2913	0.1506	1.9339	0.0531	0.0288	0.0075	1.0151	0.0073	-0.003	0.0058
European 2014	Psychiatric Sample	Mothers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.674	0.1534	4.3935	1.12E-05	0.0373	0.0079	1.0093	0.0079	-0.0131	0.0055
European 2014	Psychiatric Sample	Fathers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2446	0.1204	2.0316	0.0422	0.04	0.0069	1.0161	0.0077	0.0082	0.0055
European 2014	Psychiatric Sample	Difference in height between childhood and a	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0405	0.1174	0.3449	0.7302	0.349	0.0565	0.9663	0.0091	-0.0018	0.0056
European 2014	Psychiatric Sample	Sitting height ratio	25865494	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0998	0.1056	0.9451	0.3446	0.2263	0.0293	0.9788	0.0089	-9.76E-05	0.0054
European 2014	Psychiatric Sample	Height; Females at age 10 and males at age	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.108	0.0972	1.1112	0.2665	0.411	0.0477	0.9556	0.0092	0.0031	0.0057
European 2014	Psychiatric Sample	Sleep duration	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1531	0.0863	1.7746	0.076	0.0557	0.0051	1.0194	0.0085	-0.0042	0.0057
European 2014	Psychiatric Sample	Birth weight	27680694	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0819	0.0694	1.1806	0.2378	0.103	0.0069	1.0277	0.0096	0.0011	0.006
European 2014	Psychiatric Sample	Chronotype	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0273	0.0654	-0.4174	0.6764	0.105	0.0065	1.0045	0.0093	-0.0017	0.0054
European 2014	Psychiatric Sample	Years of schooling (proxy cognitive performar	25201988	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6344	0.0871	7.2811	3.31E-13	0.1082	0.0081	1.0216	0.0108	-7.68E-05	0.0068
European 2014	Psychiatric Sample	Years of schooling 2013	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6312	0.088	7.1739	7.29E-13	0.0865	0.0067	1.0104	0.0107	-0.0039	0.0062
European 2014	Psychiatric Sample	Neuroticism	24828478	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3201	0.1488	-2.1516	0.0314	0.0139	0.0034	1.0115	0.0073	0.0029	0.0057
European 2014	Psychiatric Sample	Age of first birth	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5519	0.0777	7.1017	1.23E-12	0.0618	0.0038	0.9534	0.0096	-0.0008	0.006
European 2014	Psychiatric Sample	Number of children ever born	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2602	0.0896	-2.9047	0.0037	0.0255	0.0019	0.9697	0.008	-0.0011	0.0056
European 2014	Psychiatric Sample	Attention deficit hyperactivity disorder (GC)	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5701	0.2338	-2.438	0.0148	0.0719	0.0295	0.9921	0.0081	0.0055	0.0062
European 2014	Psychiatric Sample	Attention deficit hyperactivity disorder (No G	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5708	0.233	-2.4499	0.0143	0.0732	0.0299	1.0066	0.0082	0.0056	0.0063
European 2014	Psychiatric Sample	Intelligence	28530673	cognitive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4638	0.0818	5.6703	1.43E-08	0.1977	0.0113	0.9937	0.0098	0.0047	0.0058
European 2014	Psychiatric Sample	Insomnia	28604731	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1524	0.0975	-1.5634	0.118	0.0485	0.0052	0.9997	0.0079	-0.0073	0.0054
European 2014	Psychiatric Sample	Insomnia	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1536	0.0944	-1.6272	0.1037	0.1372	0.0129	0.9972	0.0091	-0.0013	0.0059
European 2014	Psychiatric Sample	Excessive daytime sleepiness	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0512	0.0955	0.5358	0.5921	0.0508	0.0052	1.0116	0.0082	-0.001	0.0052
European 2014	Psychiatric Sample	Cigarettes Per Day	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.2866	0.1663	-1.7233	0.0848	0.0153	0.0081	1.0023	0.0144	0.0056	0.0057
European 2014	Psychiatric Sample	Smoking Cessation	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.306	0.2486	1.2309	0.2184	0.0087	0.0033	1.0168	0.0233	0.0078	0.017
European 2014	Psychiatric Sample	Smoking Initiation	30617275	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4074	0.1673	-2.4346	0.0149	0.0036	0.0011	1.0904	0.0107	-0.0049	0.007
European 2014	Psychiatric Sample	Pack Years	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.4483	0.4107	-1.0916	0.275	0.1051	0.0451	0.9854	0.0927	0.018	0.0559
European 2014	Psychiatric Sample	Own birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.145	0.0556	2.6095	0.0091	0.0978	0.0054	1.0593	0.0149	-0.0087	0.0061
European 2014	Psychiatric Sample	Own birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.136	0.056	2.428	0.0152	0.0937	0.005	1.0481	0.0144	-0.0082	0.0062
European 2014	Psychiatric Sample	Own birth weight (fetal effect) adjusted for r	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1889	0.0819	2.3068	0.0211	0.0297	0.0029	1.0283	0.0096	-0.009	0.0054
European 2014	Psychiatric Sample	Offspring birth weight (maternal effect) adju	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0267	0.0598	0.4466	0.6552	0.1035	0.0061	1.0396	0.0118	0.0019	0.006
European 2014	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0501	0.0598	0.8377	0.4022	0.0996	0.0055	1.0236	0.0118	-0.0002	0.0059
European 2014	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0475	0.0827	-0.5742	0.5659	0.0292	0.0026	1.0192	0.0089	0.0069	0.0055

Election	Cohort	trait2	PMID	Category	ethnicity	note	rg	se	z	p	h2_obs	h2_obs_se	h2_int	h2_int_se	gcov_int	gcov_int_se
European 2014	National Representative	Age of smoking initiation	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4202	0.3021	1.3911	0.1642	0.0634	0.0185	0.9959	0.0072	-0.0133	0.0051
European 2014	National Representative	Child birth length	25281659	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2316	0.2227	1.0396	0.2985	0.1664	0.0271	0.9914	0.0091	0.0015	0.0062
European 2014	National Representative	Child birth weight	23202124	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3789	0.1966	1.9268	0.054	0.1071	0.0185	1.0035	0.0068	-0.0071	0.0048
European 2014	National Representative	Body mass index	20935630	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0862	0.0884	-0.9755	0.3293	0.1888	0.0096	0.9997	0.0125	-0.0041	0.0063
European 2014	National Representative	Body fat	26833246	anthropometric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; S	-0.0451	0.1301	-0.3462	0.7292	0.1046	0.0083	0.9034	0.0073	-0.0074	0.0054
European 2014	National Representative	Childhood IQ	23358156	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.8742	0.3214	2.7202	0.0065	0.2834	0.0473	0.9985	0.0104	-0.0036	0.0064
European 2014	National Representative	Childhood obesity	22484627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2657	0.1462	-1.8168	0.0692	0.4369	0.0418	0.9178	0.0077	0.0039	0.0055
European 2014	National Representative	Cigarettes smoked per day	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.5871	0.2813	-2.0868	0.0369	0.0528	0.0155	1.0095	0.0077	0.0029	0.0053
European 2014	National Representative	Depressive symptoms	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3187	0.1656	-1.9242	0.0543	0.0473	0.0042	1.0026	0.0094	-0.0044	0.0064
European 2014	National Representative	Years of schooling 2016	27225129	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7591	0.189	4.0165	5.91E-05	0.1261	0.0049	0.9244	0.0123	7.80E-05	0.0061
European 2014	National Representative	Extreme bmi	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1963	0.1901	-1.0324	0.3019	0.6876	0.0605	1.0281	0.0138	-0.0016	0.0086
European 2014	National Representative	Former vs Current smoker	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5491	0.2862	1.9185	0.0551	0.0595	0.0116	1.0037	0.0078	-0.0042	0.0054
European 2014	National Representative	Anorexia Nervosa	24514567	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4368	0.1853	2.3579	0.0184	0.449	0.0332	0.938	0.0086	-0.0115	0.0053
European 2014	National Representative	Extreme height	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3088	0.2468	1.2516	0.2107	1.2949	0.1154	1.0013	0.0211	-0.0156	0.0093
European 2014	National Representative	Extreme waist-to-hip ratio	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2855	0.2934	0.9733	0.3304	0.3433	0.062	0.98	0.0095	-0.017	0.0083
European 2014	National Representative	Height_2010	20881960	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0799	0.0951	0.8401	0.4009	0.2835	0.0161	0.997	0.0185	0.0011	0.007
European 2014	National Representative	Obesity class 1	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1603	0.1216	-1.3183	0.1874	0.2139	0.0117	1.013	0.0133	-0.0014	0.0073
European 2014	National Representative	Obesity class 2	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1823	0.131	-1.3917	0.164	0.1747	0.0134	1.0094	0.0128	-0.0004	0.0074
European 2014	National Representative	Obesity class 3	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1755	0.2192	-0.8009	0.4232	0.1153	0.0155	0.9826	0.0111	0.0015	0.0076
European 2014	National Representative	Overweight	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1514	0.1055	-1.4345	0.1514	0.1085	0.0066	1.0189	0.0109	-0.004	0.0062
European 2014	National Representative	Neo-conscientiousness	21173776	personality	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.1029	0.3046	-0.3377	0.7356	0.0797	0.033	0.9988	0.0086	0.0024	0.0055
European 2014	National Representative	Neo-openness to experience	21173776	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3345	0.2695	1.2413	0.2145	0.1181	0.0284	0.9885	0.0078	0.0046	0.0058
European 2014	National Representative	Hip circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0556	0.0987	-0.5632	0.5733	0.1289	0.0058	0.8519	0.0095	0.0007	0.0049
European 2014	National Representative	Alzheimers disease	24162737	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.4554	0.2498	-1.823	0.0683	0.0439	0.0196	1.0659	0.0281	0.0016	0.0053
European 2014	National Representative	Infant head circumference	22504419	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4502	0.3043	1.4797	0.1389	0.239	0.0448	0.9883	0.0077	-0.0082	0.0064
European 2014	National Representative	Age at Menarche	25231870	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0792	0.0881	-0.8988	0.3688	0.2082	0.0104	0.9294	0.0134	0.0081	0.0063
European 2014	National Representative	Age at Menopause	26414677	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0624	0.1276	0.4894	0.6245	0.1363	0.0166	0.9863	0.0135	-0.0005	0.0059
European 2014	National Representative	Neuroticism	27089181	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1109	0.1422	-0.7794	0.4358	0.0912	0.0079	0.9791	0.0149	-0.0147	0.0072
European 2014	National Representative	Attention deficit hyperactivity disorder	20732625	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.0758	0.303	-0.25	0.8026	0.2341	0.0976	1.0074	0.0077	-0.0002	0.0051
European 2014	National Representative	Bipolar disorder	21926972	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1964	0.1458	1.3471	0.178	0.4232	0.0383	1.024	0.0083	0.0034	0.0056
European 2014	National Representative	PGC cross-disorder analysis	23453885	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0669	0.1285	0.5209	0.6024	0.1747	0.0132	1.0098	0.0121	0.0028	0.0078
European 2014	National Representative	Major depressive disorder	22472876	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1197	0.1842	-0.6502	0.5155	0.1796	0.0271	1	0.0068	0.0061	0.0049
European 2014	National Representative	Autism spectrum disorder	0	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1002	0.1654	0.606	0.5445	0.4992	0.0565	0.9565	0.0077	0.0095	0.0054
European 2014	National Representative	Parkinsons disease	19915575	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0646	0.177	0.365	0.7151	0.3178	0.1354	1.13	0.0114	-0.0006	0.005
European 2014	National Representative	Schizophrenia	25056061	psychiatric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; S	-0.1153	0.1011	-1.1407	0.254	0.4615	0.0196	1.0464	0.0148	-0.0019	0.0075
European 2014	National Representative	College completion	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7555	0.2381	3.1732	0.0015	0.0788	0.0062	1.0202	0.0098	0.0083	0.0055
European 2014	National Representative	Subjective well being	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2519	0.1779	1.4164	0.1567	0.0244	0.0023	1.003	0.0089	0.0014	0.006
European 2014	National Representative	Ever vs never smoked	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.528	0.2077	-2.5417	0.011	0.0701	0.0069	1.0026	0.0073	-4.31E-06	0.0056
European 2014	National Representative	Waist circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0892	0.0961	-0.9279	0.3535	0.1224	0.0052	0.8373	0.0094	-0.0038	0.0054
European 2014	National Representative	Waist-to-hip ratio	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0967	0.097	-0.9968	0.3188	0.1116	0.0073	0.9146	0.0106	-0.0067	0.0058
European 2014	National Representative	Amyotrophic lateral sclerosis	27455348	neurological	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.3514	0.2341	-1.501	0.1334	0.0568	0.0135	0.9874	0.0069	0.0047	0.0051
European 2014	National Representative	Difference in height between adolescence a	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3556	0.228	-1.56	0.1188	0.4873	0.1046	0.9757	0.007	0.009	0.0051
European 2014	National Representative	Parents age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7649	0.2584	2.9602	0.0031	0.0288	0.0076	1.0151	0.0073	-0.0147	0.0053
European 2014	National Representative	Mothers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.799	0.2809	2.8445	0.0044	0.0378	0.0079	1.0085	0.008	-0.0134	0.0049
European 2014	National Representative	Fathers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5026	0.2004	2.5083	0.0121	0.0401	0.0069	1.0159	0.0077	-0.0029	0.0057
European 2014	National Representative	Difference in height between childhood and	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1501	0.1937	-0.7749	0.4384	0.3489	0.0567	0.9663	0.0091	0.0019	0.0062
European 2014	National Representative	Sitting height ratio	25865494	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1162	0.142	0.8184	0.4131	0.2271	0.0292	0.9784	0.0088	-0.0026	0.0056
European 2014	National Representative	Height; Females at age 10 and males at age	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.3363	0.1839	1.8285	0.0675	0.4108	0.0477	0.9557	0.0092	-0.0047	0.0059
European 2014	National Representative	Sleep duration	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2734	0.1564	1.7483	0.0804	0.0561	0.005	1.0185	0.0085	-0.0093	0.0052
European 2014	National Representative	Birth weight	27680694	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2507	0.1153	2.1744	0.0297	0.1028	0.0069	1.0281	0.0096	-0.0055	0.0057
European 2014	National Representative	Chronotype	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1197	0.1161	1.0311	0.3025	0.1052	0.0065	1.004	0.0093	-0.0072	0.0054
European 2014	National Representative	Years of schooling (proxy cognitive perform	25201988	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6949	0.2296	3.0273	0.0025	0.1082	0.0081	1.0216	0.0108	0.0129	0.006
European 2014	National Representative	Years of schooling 2013	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7008	0.2049	3.4206	0.0006	0.0865	0.0067	1.0105	0.0107	0.011	0.0058
European 2014	National Representative	Neuroticism	24828478	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0178	0.2429	-0.0734	0.9415	0.0139	0.0034	1.0115	0.0073	-0.0089	0.0056
European 2014	National Representative	Age of first birth	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6578	0.1822	3.6101	0.0003	0.0617	0.0038	0.9538	0.0097	0.002	0.0063
European 2014	National Representative	Number of children ever born	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1231	0.1429	-0.8616	0.3889	0.0254	0.0019	0.9703	0.008	-0.0072	0.0054
European 2014	National Representative	Attention deficit hyperactivity disorder (GC)	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5733	0.3966	-1.4457	0.1483	0.0719	0.0295	0.9921	0.0081	0.0006	0.0061
European 2014	National Representative	Attention deficit hyperactivity disorder (No	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5764	0.399	-1.4445	0.1486	0.0732	0.0299	1.0066	0.0082	0.0007	0.0062
European 2014	National Representative	Intelligence	28530673	cognitive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.687	0.179	3.8375	0.0001	0.1984	0.0113	0.9928	0.0098	-0.0112	0.0057
European 2014	National Representative	Insomnia	28604731	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3813	0.1731	-2.2035	0.0276	0.0485	0.0052	0.9995	0.0079	0.0057	0.0058
European 2014	National Representative	Insomnia	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3095	0.1483	-2.0869	0.0369	0.1371	0.0129	0.9974	0.0091	0.004	0.0059
European 2014	National Representative	Excessive daytime sleepiness	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2124	0.1586	-1.3393	0.1805	0.051	0.0052	1.0112	0.0081	0.0087	0.0058
European 2014	National Representative	Cigarettes Per Day	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.5231	0.292	-1.7915	0.0732	0.0156	0.0081	1.0012	0.0143	0.0107	0.0065
European 2014	National Representative	Smoking Cessation	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.2902	0.3988	0.7276	0.4668	0.0089	0.0033	1.0159	0.0228	0.0042	0.019
European 2014	National Representative	Smoking Initiation	30617275	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.5063	0.3068	-1.6502	0.0989	0.0036	0.0011	1.0911	0.0107	0.0004	0.0071
European 2014	National Representative	Pack Years	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.8548	0.5523	-1.5477	0.1217	0.1165	0.0468	0.9604	0.0949	0.1284	0.078
European 2014	National Representative	Own birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2622	0.0976	2.6869	0.0072	0.0976	0.0054	1.0601	0.015	-0.0132	0.0065
European 2014	National Representative	Own birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2537	0.0929	2.729	0.0064	0.0936	0.0051	1.0485	0.0145	-0.0136	0.0064
European 2014	National Representative	Own birth weight (fetal effect) adjusted for	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1915	0.1253	1.5284	0.1264	0.0297	0.0029	1.0283	0.0096	-0.0035	0.0058
European 2014	National Representative	Offspring birth weight (maternal effect) adj	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1932	0.1013	1.9063	0.0566	0.1035	0.0061	1.0399	0.0117	-0.0141	0.0062
European 2014	National Representative	Offspring birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.167	0.0999	1.6717	0.0946	0.0996	0.0055	1.0239	0.0118	-0.0139	0.0063
European 2014	National Representative	Offspring birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2228	0.132	1.6885	0.0913	0.0292	0.0026	1.0192	0.0088	-0.0111	0.0059

Election	Cohort	trait2	PMID	Category	ethnicity	note	rg	se	z	p	h2_obs	h2_obs_se	h2_int	h2_int_se	gcov_int	gcov_int_se
National 2015	Psychiatric Sample	Age of smoking initiation	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2294	0.2074	-1.1059	0.2688	0.0644	0.0185	0.9955	0.0072	0.008	0.0054
National 2015	Psychiatric Sample	Child birth length	25281659	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1929	0.1554	1.2411	0.2146	0.1664	0.0271	0.9914	0.0091	-0.0005	0.0061
National 2015	Psychiatric Sample	Child birth weight	23202124	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.181	0.138	1.3112	0.1898	0.1082	0.0185	1.0029	0.0068	0.0004	0.0048
National 2015	Psychiatric Sample	Body mass index	20935630	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1244	0.0812	-1.5331	0.1252	0.1888	0.0096	0.9997	0.0125	-0.0046	0.0067
National 2015	Psychiatric Sample	Body fat	26833246	anthropometric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; SI	-0.252	0.1005	-2.508	0.0121	0.1043	0.0083	0.9038	0.0074	0.005	0.0051
National 2015	Psychiatric Sample	Childhood IQ	23358156	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.7908	0.1857	4.2596	2.05E-05	0.283	0.0473	0.9986	0.0104	-0.0119	0.0062
National 2015	Psychiatric Sample	Childhood obesity	22484627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0659	0.1215	-0.5426	0.5874	0.4355	0.0419	0.9183	0.0078	-0.0007	0.0057
National 2015	Psychiatric Sample	Cigarettes smoked per day	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3067	0.1824	-1.6813	0.0927	0.0523	0.0154	1.01	0.0077	-0.0011	0.0051
National 2015	Psychiatric Sample	Depressive symptoms	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2778	0.1142	-2.4327	0.015	0.0473	0.0042	1.0026	0.0094	-0.0063	0.0063
National 2015	Psychiatric Sample	Years of schooling 2016	27225129	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6172	0.0814	7.5784	3.50E-14	0.1261	0.0049	0.9247	0.0123	-0.0055	0.0064
National 2015	Psychiatric Sample	Extreme bmi	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1585	0.1283	-1.2353	0.2167	0.6876	0.0605	1.0281	0.0138	-0.0002	0.009
National 2015	Psychiatric Sample	Former vs Current smoker	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4206	0.1519	2.7693	0.0056	0.0598	0.0116	1.0035	0.0078	-0.004	0.005
National 2015	Psychiatric Sample	Anorexia Nervosa	24514567	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0704	0.1033	0.6816	0.4955	0.4493	0.0332	0.9379	0.0086	-0.0018	0.0065
National 2015	Psychiatric Sample	Extreme height	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0344	0.0894	0.3848	0.7004	1.2949	0.1154	1.0013	0.0211	0.0126	0.0091
National 2015	Psychiatric Sample	Extreme waist-to-hip ratio	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1146	0.1656	-0.6918	0.4891	0.3433	0.062	0.98	0.0095	-0.0052	0.0077
National 2015	Psychiatric Sample	Height_2010	20881960	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0298	0.0616	0.4835	0.6288	0.2839	0.0161	0.996	0.0186	0.0086	0.0067
National 2015	Psychiatric Sample	Obesity class 1	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1698	0.0909	-1.8674	0.0618	0.2139	0.0117	1.013	0.0133	-0.0113	0.0075
National 2015	Psychiatric Sample	Obesity class 2	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1737	0.1229	-1.4135	0.1575	0.1747	0.0134	1.0094	0.0128	-0.009	0.0091
National 2015	Psychiatric Sample	Obesity class 3	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2983	0.1496	-1.9946	0.0461	0.1153	0.0155	0.9826	0.0111	0.0028	0.0078
National 2015	Psychiatric Sample	Overweight	23563607	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1462	0.0818	-1.7859	0.0741	0.1086	0.0066	1.0189	0.0109	-0.0021	0.0069
National 2015	Psychiatric Sample	Neo-conscientiousness	21173776	personality	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.1869	0.2456	0.7613	0.4465	0.08	0.033	0.9987	0.0086	-0.0021	0.0064
National 2015	Psychiatric Sample	Neo-openness to experience	21173776	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4173	0.1683	2.4792	0.0132	0.1181	0.0284	0.9885	0.0077	-0.0036	0.0054
National 2015	Psychiatric Sample	Hip circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0942	0.0741	-1.2719	0.2034	0.1287	0.0059	0.8523	0.0095	0.0057	0.0059
National 2015	Psychiatric Sample	Alzheimers disease	24162737	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0766	0.1642	-0.4667	0.6407	0.0439	0.0195	1.0659	0.0281	-0.0045	0.006
National 2015	Psychiatric Sample	Infant head circumference	22504419	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0685	0.1596	-0.429	0.6679	0.2391	0.0449	0.9883	0.0078	0.0047	0.006
National 2015	Psychiatric Sample	Age at Menarche	25231870	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0253	0.0687	-0.3675	0.7132	0.2084	0.0104	0.9289	0.0134	0.0022	0.0059
National 2015	Psychiatric Sample	Age at Menopause	26414677	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2166	0.0965	2.2458	0.0247	0.1367	0.0166	0.9857	0.0135	-0.0075	0.006
National 2015	Psychiatric Sample	Neuroticism	27089181	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1215	0.0932	-1.3033	0.1925	0.0912	0.0079	0.9791	0.0149	-0.0103	0.0067
National 2015	Psychiatric Sample	Attention deficit hyperactivity disorder	20732625	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.024	0.2228	0.1076	0.9143	0.2342	0.0974	1.0073	0.0077	-0.0091	0.0055
National 2015	Psychiatric Sample	Bipolar disorder	21926972	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1132	0.1223	0.9257	0.3546	0.425	0.0383	1.0234	0.0083	0.0026	0.0062
National 2015	Psychiatric Sample	PGC cross-disorder analysis	23453885	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0146	0.1053	0.1383	0.89	0.1754	0.0132	1.0089	0.012	-0.003	0.0078
National 2015	Psychiatric Sample	Major depressive disorder	22472876	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1031	0.1354	-0.7618	0.4462	0.1803	0.0271	0.9997	0.0068	-0.0009	0.0059
National 2015	Psychiatric Sample	Autism spectrum disorder	0	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.326	0.1195	2.7283	0.0064	0.4974	0.0565	0.9568	0.0077	-0.0113	0.0054
National 2015	Psychiatric Sample	Parkinsons disease	19915575	neurological	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1879	0.1619	-1.1606	0.2458	0.3178	0.1354	1.13	0.0114	0.0082	0.0049
National 2015	Psychiatric Sample	Schizophrenia	25056061	psychiatric	Mixed	Caution: using this data may yield less robust results due to minor departure of the LD structure; SI	-0.1615	0.0671	-2.4075	0.0161	0.4614	0.0196	1.0464	0.0148	-0.0067	0.0077
National 2015	Psychiatric Sample	College completion	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6261	0.119	5.2596	1.44E-07	0.0788	0.0062	1.0202	0.0098	-0.0037	0.006
National 2015	Psychiatric Sample	Subjective well being	27089181	psychiatric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1651	0.1158	1.4259	0.1539	0.0244	0.0023	1.003	0.0089	0.005	0.0067
National 2015	Psychiatric Sample	Ever vs never smoked	20418890	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1907	0.1172	-1.6268	0.1038	0.07	0.0069	1.0029	0.0072	-0.0024	0.0053
National 2015	Psychiatric Sample	Waist circumference	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.124	0.0724	-1.7124	0.0868	0.1222	0.0053	0.8378	0.0094	-0.0015	0.0056
National 2015	Psychiatric Sample	Waist-to-hip ratio	25673412	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.1288	0.07	-1.8407	0.0657	0.1115	0.0073	0.9149	0.0106	-0.0053	0.0054
National 2015	Psychiatric Sample	Amyotrophic lateral sclerosis	27455348	neurological	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.3742	0.164	-2.282	0.0225	0.0568	0.0135	0.9874	0.0069	0.003	0.0052
National 2015	Psychiatric Sample	Difference in height between adolescence a	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.081	0.1582	-0.5119	0.6088	0.4889	0.1052	0.9755	0.0071	0.0021	0.005
National 2015	Psychiatric Sample	Parents age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4684	0.1647	2.8441	0.0045	0.0288	0.0075	1.015	0.0073	-0.0042	0.0054
National 2015	Psychiatric Sample	Mothers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.5811	0.1562	3.7199	0.0002	0.0374	0.0079	1.0092	0.008	-0.0044	0.0056
National 2015	Psychiatric Sample	Fathers age at death	27015805	aging	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4594	0.1482	3.1003	0.0019	0.04	0.0069	1.016	0.0078	-0.0055	0.0057
National 2015	Psychiatric Sample	Difference in height between childhood and	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.2479	0.147	1.6868	0.0916	0.3488	0.0566	0.9664	0.0091	-0.006	0.0056
National 2015	Psychiatric Sample	Sitting height ratio	25865494	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.1137	0.1366	0.8317	0.4056	0.2269	0.0292	0.9786	0.0089	-0.0052	0.006
National 2015	Psychiatric Sample	Height; Females at age 10 and males at age	23449627	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0255	0.1115	0.2286	0.8192	0.4111	0.0477	0.9555	0.0092	0.0051	0.0057
National 2015	Psychiatric Sample	Sleep duration	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0271	0.0985	0.2751	0.7833	0.0558	0.0051	1.0192	0.0085	0.0048	0.0054
National 2015	Psychiatric Sample	Birth weight	27680694	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0656	0.0765	0.8571	0.3914	0.1031	0.0069	1.0275	0.0097	0.0008	0.006
National 2015	Psychiatric Sample	Chronotype	27494321	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0437	0.0828	-0.5275	0.5978	0.1053	0.0065	1.0038	0.0093	-0.0048	0.0056
National 2015	Psychiatric Sample	Years of schooling (proxy cognitive performa	25201988	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6612	0.1138	5.8086	6.30E-09	0.1082	0.0081	1.0216	0.0108	-0.0023	0.0066
National 2015	Psychiatric Sample	Years of schooling 2013	23722424	education	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6222	0.1087	5.7233	1.04E-08	0.0865	0.0067	1.0104	0.0107	-0.0007	0.0063
National 2015	Psychiatric Sample	Neuroticism	24828478	personality	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2394	0.1838	-1.3028	0.1926	0.0139	0.0034	1.0115	0.0073	1.29E-06	0.0062
National 2015	Psychiatric Sample	Age of first birth	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.6145	0.0922	6.6663	2.62E-11	0.0619	0.0038	0.9531	0.0097	-0.0074	0.006
National 2015	Psychiatric Sample	Number of children ever born	27798627	reproductive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2566	0.1016	-2.5249	0.0116	0.0255	0.0019	0.9698	0.008	0.0066	0.0059
National 2015	Psychiatric Sample	Attention deficit hyperactivity disorder (GC)	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.7982	0.3263	-2.4463	0.0144	0.0719	0.0295	0.9921	0.0081	0.0081	0.0064
National 2015	Psychiatric Sample	Attention deficit hyperactivity disorder (No (	27663945	psychiatric	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.7937	0.3247	-2.4446	0.0145	0.0732	0.0299	1.0066	0.0082	0.0082	0.0065
National 2015	Psychiatric Sample	Intelligence	28530673	cognitive	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.4935	0.0859	5.745	9.19E-09	0.1983	0.0112	0.9929	0.0098	-0.0079	0.0056
National 2015	Psychiatric Sample	Insomnia	28604731	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2168	0.1253	-1.7297	0.0837	0.0485	0.0052	0.9996	0.0079	-0.005	0.0059
National 2015	Psychiatric Sample	Insomnia	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.2958	0.1096	-2.698	0.007	0.1374	0.0128	0.997	0.0091	-0.0021	0.006
National 2015	Psychiatric Sample	Excessive daytime sleepiness	27992416	sleeping	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.0644	0.1063	-0.6057	0.5447	0.0509	0.0052	1.0114	0.0082	6.82E-05	0.0057
National 2015	Psychiatric Sample	Cigarettes Per Day	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.003	0.1388	-0.0219	0.9825	0.0155	0.0081	1.0018	0.0143	-0.0082	0.0058
National 2015	Psychiatric Sample	Smoking Cessation	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	0.3439	0.3663	0.9388	0.3478	0.0087	0.0033	1.0167	0.0234	0.0058	0.0187
National 2015	Psychiatric Sample	Smoking Initiation	30617275	smoking_behaviour	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	-0.3288	0.1807	-1.8195	0.0688	0.0037	0.0011	1.0897	0.0106	-0.0074	0.0064
National 2015	Psychiatric Sample	Pack Years	30617275	smoking_behaviour	European	Caution: using this data may yield results outside bounds due to relative low Z score of the SNP he	-0.4581	0.29	-1.5796	0.1142	0.112	0.0449	0.9697	0.0892	0.0353	0.0562
National 2015	Psychiatric Sample	Own birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0663	0.0608	1.0921	0.2748	0.0978	0.0054	1.0594	0.0149	-0.0022	0.006
National 2015	Psychiatric Sample	Own birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0686	0.0592	1.1586	0.2466	0.0937	0.0051	1.0483	0.0144	-0.0025	0.0056
National 2015	Psychiatric Sample	Own birth weight (fetal effect) adjusted for	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0423	0.0894	0.4732	0.636	0.0298	0.0029	1.0281	0.0096	0.0007	0.0055
National 2015	Psychiatric Sample	Offspring birth weight (maternal effect) adj	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0276	0.0661	0.4175	0.6763	0.1035	0.0061	1.0398	0.0118	-0.0009	0.0056
National 2015	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	Mixed	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0206	0.0641	0.321	0.7482	0.0995	0.0055	1.0241	0.0118	-0.0002	0.0057
National 2015	Psychiatric Sample	Offspring birth weight	31043758	anthropometric	European	SNPs from the MHC (chr6 26M~34M) region was removed for this traits	0.0315	0.0914	0.345	0.7301	0.0292	0.0026	1.0192	0.0089	-0.0005	0.0054

DECLARATION OF CO-AUTHORSHIP

The declaration is for PhD students and must be completed for each conjointly authored article. Please note that if a manuscript or published paper has ten or less co-authors, all co-authors must sign the declaration of co-authorship. If it has more than ten co-authors, declarations of co-authorship from the corresponding author(s), the senior author and the principal supervisor (if relevant) are a minimum requirement.

1. Declaration by	
Name of PhD student	Vivek Appadurai
E-mail	vivek.appadurai@regionh.dk
Name of principal supervisor	Dr. Thomas Werge
Title of the PhD thesis	Genetic Analysis of Complex Traits in Population Scale Datasets

2. The declaration applies to the following article	
Title of article	Genetic predictors of educational attainment and intelligence test performance predict voter turnout
Article status	
Published ☐ Date:	Accepted for publication ■ Date: 17-08-2020
Manuscript submitted ☐ Date:	Manuscript not submitted ☐
If the article is published or accepted for publication, please state the name of journal, year, volume, page and DOI (if you have the information).	Nature Human Behaviour https://doi.org/10.1038/s41562-020-00952-2

3. The PhD student's contribution to the article *(please use the scale A-F as benchmark)* Benchmark scale of the PhD-student's contribution to the article **A.** Has essentially done all the work (> 90 %) **B.** Has done most of the work (60-90 %) **C.** Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
1. Formulation/identification of the scientific problem	C
2. Development of the key methods	B
3. Planning of the experiments and methodology design and development	B
4. Conducting the experimental work/clinical studies/data collection/obtaining access to data	C

3. The PhD student's contribution to the article *(please use the scale A-F as benchmark)* Benchmark scale of the PhD-student's contribution to the article **A.** Has essentially done all the work (> 90 %) **B.** Has done most of the work (60-90 %) **C.** Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
5. Conducting the analysis of data	A
6. Interpretation of the results	B
7. Writing of the first draft of the manuscript	C
8. Finalisation of the manuscript and submission	C

Provide a short description of the PhD student´s specific contribution to the article.[i]

Contributed to the study design, analysis plan and quality control of the data. Performed phasing and imputation of genotypes, performed data analysis including estimates of heritability, genome wide association studies, calculation of genetic correlations, polygenic score analysis and mendelian randomization. Wrote the methods section of the manuscript, contributed to the final draft.

4. Material from another thesis / dissertation[ii]	
Does the article contain work which has also formed part of another thesis, e.g. master's thesis, PhD thesis or doctoral dissertation (the PhD student's or another person's)?	Yes: ☐ No: ■
If yes, please state name of the author and title of thesis / dissertation.	
If the article is part of another author's academic degree, please describe the PhD student´s and the author´s contributions to the article so that the individual contributions are clearly distinguishable from one another.	

5. Signatures of the co-authors[iii]

	Date	Name	Title	Signature
1.		Lene Aaroe	PhD	Lene Aaroe
2.		Andrew J Schork	PhD	Andrew Schork

5. Signatures of the co-authors[iii]				
3.	14/12-2020	Thomas Werge	PhD	
4.		Alfonso Buil	PhD	
5.		Esben Agerbo	PhD	
6.		Michael Bang Petersen	PhD	Michael Bang Petersen
7.				
8.				
9.				
10.				

6. Signature of the principal supervisor

I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge.

Date: 14/12-2020

Principal supervisor:

7. Signature of the PhD student

I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge.

Date: 10-12-2020

PhD student: VIVEK APPADURAI

Please learn more about responsible conduct of research on the Faculty of Health and Medical Sciences' website.

[i] This can be supplemented with an additional letter if needed.

[ii] Please see Ministerial Order on the PhD Programme at the Universities and Certain Higher Artistic Educational Institutions (PhD Order) § 12 (4):

"Any articles included in the thesis may be written in cooperation with others, provided that each of the co-authors submits a written declaration stating the PhD student's or the author's contribution to the work."

[iii] If more signatures are needed please add an extra sheet.

PAPER 3

Appadurai V, Rosengren A, Bybjerg-Grauholm J, Ingason A, Buil A, Mors O, Børglum AD, Hougaard DM, Nordentoft M, Mortensen PB, Werge T, Delaneau O, Schork AJ. **Legacy data, whole genome imputation and the analysis of complex traits: Lessons from the iPSYCH case-cohort study**. Manuscript in preparation.

MANUSCRIPT

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TABLES

DECLARATION OF CO-AUTHORSHIP

Legacy data, whole genome imputation, and the analysis of complex traits: Lessons from the iPSYCH case-cohort study

Vivek Appadurai[1,2], Anders Rosengren[1,2], Jonas Grauholm[2,3], Alfonso Buil[1,2], Andrés Ingason[1,2], Ole Mors[2,6], Anders D. Børglum[2,7], David M. Hougaard[3,8], Merete Nordentoft[2,9], Preben B. Mortensen[2,10,11], Thomas Werge[1,2], Olivier Delaneau[5], Andrew J. Schork[1,2,4]

1. *Institute of Biological Psychiatry, Mental Health Center Sankt Hans, Roskilde, Denmark 4000*
2. *The Lundbeck Initiative for Integrated Psychiatric Research, Aarhus, Denmark*
3. *Danish Center for Neonatal Screening, Statens Serum Institut, Copenhagen, Denmark*
4. *The Translational Genomics Research Institute, Phoenix AZ, USA*
5. *Department of Computational Biology, University of Lausanne, Switzerland*
6. *Psychosis Research Unit, Aarhus University Hospital - Psychiatry, Aarhus, Denmark*
7. *Department of Biomedicine and Center for Integrative Sequencing, iSEQ, Aarhus University, Denmark*
8. *Center for Genomics and Personalized Medicine, Aarhus University, Aarhus, Denmark*
9. *Mental Health Centre, Copenhagen, Capital Region of Denmark, Copenhagen, Denmark*
10. *NCRR - National Center for Register-Based Research, Business and Social Sciences, Aarhus University, Aarhus, Denmark*
11. *CIRRAU - Centre for Integrated Register-Based Research, Aarhus University, Aarhus, Denmark*

Abstract

Haplotype estimation (phasing) followed by whole genome imputation (imputation) is a bedrock of complex trait genetic analysis. The sample recruitments for international research consortia, hospital systems, national biobanks and direct to consumer DNA testing companies often span multiple years and necessitate genotyping in several stages, using different technologies at each stage. The integration of legacy data is non-trivial, as the coverage and quality of array data may vary systematically and the quality of resulting phasing and imputation have not been systematically evaluated. The Lundbeck foundation initiative for integrative psychiatric research (iPSYCH) consortium dataset comprises 130,438 individuals, genotyped in two stages, each comprising multiple batches, with two different genotyping arrays. Here we evaluated the accuracy of haplotype estimation and whole genome imputation using three different tools across four different data integration protocols. Phasing accuracy as measured by switch error rate varied by both the choice of tool and data integration protocol. The empirical correlation between true genotypes and imputed allele dosages varied more by choice of data integration protocol than by choice of phasing tool. The haplotypes estimated using Beagle5 provide consistently better imputation accuracy across all data integration protocols. In addition, we show a substantial attenuation in the accuracy of imputation within samples of non-European origin, even when using the latest tools with the haplotype reference consortium v1.1 (HRCv1.1) as reference panel and even when restricting to common variants. An attenuation between 5.2% and 13.2% is observed in variance explained by simulated polygenic scores when using imputed data instead of true genotypes. Comparing random population samples genotyped and imputed using different technologies suggested protocol specific bath effects where present in each data set.

Introduction

Owing to the recent appreciation for the polygenic nature of complex traits, with risk loci scattered throughout the genome conferring small individual effects, sample sizes in the hundreds of thousands are more less required genome wide association studies[1,2]. Due to their cost effective nature, genotyping arrays, which typically ascertain single nucleotide polymorphisms (SNP) at anywhere between 200,000 and 2 million SNPs in the human genome, have become the preferred technology for generating genetic data at such sample sizes and a key component of these studies is genotype imputation[3].

Genotype imputation can be roughly thought of as a two-step process. First, a collection of genotyped SNPs are organized into haplotype scaffolds. This process uses inferred statistics describing the co-inheritance pattern of SNPs to *phase* these unlinked, individual genotypes into the linked multi-allele segments from a common ancestor, the haplotype scaffolds. These haplotype scaffolds can then be probabilistically imputed at known, untyped variants by matching the sparse scaffolds to more densely constructed reference haplotypes from, for example, whole genome sequenced individuals[4]. This process results in a much larger pool of variants that can increase the power of association studies[5], provide a set of SNPs for meta-analysis across cohorts genotyped on different arrays[6], and ensure sufficient overlap of SNPs between reference and target datasets for polygenic scoring[7]. However, there are numerous tools for performing this process, multiple reference data sets that are employed, research cohorts beginning with different marker sets, and diverse batch or legacy data is often combined, even within a single cohort.

State of the field methods such as BEAGLE5[8], SHAPEIT4[9] and EAGLE2[10] use hidden markov model approaches built on the Li and Stephens model[11]. This model assumes each individual's haplotype can be estimated as a mosaic of segments from haplotypes observed in reference data and/or the study population, modeling additional factors such as recombination and mutation rates in the population. Current tools differ in the computational approximations and data structures used for storing the most informative haplotypes for phasing each segment of the target dataset. Each tool further gives the user different parameters to increase the number of informative haplotypes considered, with a trade-off between accuracy, run times, and memory usage. While these tools have been improved over the years to scale computationally with large genetic datasets such as the UK biobank[12], the benchmarking is often performed in the same datasets, such as subsets of the 1000 genomes project samples[13], the UK biobank, the genome in a bottle dataset[14] , or the GERA cohort[15].

To the best of our knowledge, the robustness of these methods has not been characterized in input datasets with varying SNP density, target sample sizes, and missingness that can arise when integrating data generated from multiple or diverse genotyping platforms. It is important to empirically characterize the accuracy of phasing and imputation in such scenarios so that researchers can make educated choices when designing modern bioinformatics workflows.

The predominant approach used by complex trait analysis research consortia for analyzing samples genotyped on multiple arrays has been to phase and impute them separately prior to meta-analyzing the results for genome wide association studies[16,17]. However, the accuracy of phasing has been demonstrated to increase with increased sample sizes of the reference *and* target datasets. Moreover for samples generated from recent population-scale biobanks (e.g., UKbiobank[12], iPSYCH[18]), the number of study individuals is often much greater than the largest available haplotype reference data set. Cryptic relatedness among study individuals and geographical variation in haplotype frequency suggest these study haplotypes are likely to be at least as informative as published references for improving quality of phased haplotype scaffolds [19]. Hence, there is an intuitive reasoning for pooling together as many samples as possible for haplotype scaffold estimation, prior to carrying out imputation. In the UK Biobank study, where 500,000 participants were genotyped in 33 batches using two genotyping arrays, it was possible to phase and impute the entire study population together, leveraging the unprecedented sample size, as the arrays used, the UK Biobank Axiom array and the UK BiLEVE array were closely matched with a 95% marker overlap (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf). However, challenges arise in scenarios where the data integration process involves different arrays with low marker overlap and there is insufficient guiding research at present.

Earlier studies on integrating cohorts genotyped on different arrays were on a much smaller scale than current applications, used prior iterations of tools, or focused on less diverse data samples. Sinnott et. al (2012)[20] compared the frequencies of imputed alleles in two groups of European ancestry healthy controls that were genotyped on different arrays with only ~30% overlap. They found nearly 10,000 out of the 2.3 million imputed SNPs were associated at genome-wide significance with the array that the samples were imputed on - a substantial type I error rate. Retaining only the set of SNPs imputed at the highest quality reduced but did not completely eliminate these errors. Uh et. al (2012)[21] combined two data sets imputed from arrays that overlapped at only 60,000 markers. GWAS across all markers deemed good quality showed an inflation in test statistics that was much higher than when restricting to the markers genotyped on both arrays or only including subjects genotyped

on one array. The inflation was reduced when an extreme quality control was applied (r^2 quality metric > 0.98). Johnson et. al (2013)[22] compared two approaches for integrating cases and controls genotyped on different arrays. They observed that imputing from the union of SNPs across arrays led to 0.2% of SNPs showing associations to genotyping array while imputing from the intersection of the SNPs on both arrays led to lower estimated imputation accuracy, but without the same array bias. These previous studies highlight challenges associated with integrating legacy genotype data, but do not provide a clear consensus as to a direction forward. Technical artifacts arising during the genotype data generation process is one of the causes for lack of replicability of association signals from GWAS and the loss of predictive performance from polygenic scores between different cohorts[23]. Since samples collected through biobank initiatives have evolved to be the largest contributing cohorts towards GWAS conducted by large consortia[24,25], it is vital to identify and account for the artifacts that can arise during the data generation process in such large, population scale datasets where legacy data integration becomes necessary.

Beyond association studies, Pimental et. al[26] studied the biases introduced by imputation in context of estimating direct genomic values (DGV) in livestock, an analog to polygenic scores in the current parlance of complex trait human genetics. They observed an apparent bias in imputed genotypes towards the more frequent (major) allele in the reference panel that caused estimated DGVs to be shrunk towards the sample mean. This bias was more evident in traits with high heritability when DGVs were estimated using imputation from low density markers. From our experience, most discussions around imputation accuracy are centered around the correlation between true and imputed genotypes. While this is a key statistic for considering the subsequent power of GWAS, it may ignore the possible effects of directional bias and under value error in estimation problems like DGV and polygenic scores. Given the attention polygenic scores have recently received[23,24,27,28], exploring these concepts in modern, population-scale, human complex trait genetics applications is critical.

This study uses the iPSYCH case-cohort dataset with 81,330 subjects genotyped on the Infinium PsychChip v1.0 (Illumina, San Diego, CA USA) and 49,108 subjects genotyped on the Illumina Global Screening Array v2.0 (Illumina San Diego, CA USA) to evaluate four realistic protocols for data integration. We compare the accuracy of haplotype estimation using SHAPEIT4.1.2, EAGLE2.4.1, BEAGLE5, and a consensus approach using truth sets derived from 124 parent-offspring trios genotyped on both iPSYCH arrays. To compare the resulting imputation quality, we masked 10,000 SNPs prior to phasing and included 10 whole genome sequenced samples from the Personal Genomes Project - UK cohort[29], down-sampled to each phasing SNP set. The BEAGLE5.1 imputed genotypes was

then compared to these truth sets, assessing both a loss of information and direction bias in estimated dosage data. It is known that current reference haplotype data sets are heavily skewed towards individuals of European ancestry, as so we assessed quality of phasing and imputation in subsets of individuals with non-European and admixed genetic ancestries. Finally, we explored the impact of using imputed genotypes for polygenic scores using simulation studies.

Methods and Materials

Data

iPSYCH2012 is a case-cohort design nested within 1,472,762 individuals born in Denmark in the time period spanning the 1st of May 1981 and the 31st of December 2005, with a known mother, alive and residing in Denmark at the end of their first birth year. Out of 86,189 individuals chosen for genotyping, 57,377 individuals constitute cases with one or more mental disorders including schizophrenia, autism, attention-deficit/hyperactivity disorder and affective disorder. The cohort is a random sample of 30,000 individuals representative of the national population of Denmark born in the same time period. Genotyping was performed using DNA from dried blood spots obtained from the Danish neonatal screening biobank[30] at The Broad Institute, Boston MA, USA with the Infinium PsychChip Array v1.0 Illumina, San Diego CA, USA). Further details on the ascertainment and data generation process of iPSYCH2012 has previously been described[18].

iPSYCH2015i is an extension of iPSYCH2012 and it is nested within 1,717,316 individuals born in Denmark in the time spanning 1st of May 1981 and 31st of December 2008 with a known mother and alive, residing in Denmark at the end of the first birth year. The sample ascertained for genotyping consists of 33,345 cases with one or more severe mental disorders and a further 15,756 individuals constituting the cohort as a random representative population of Denmark from the same time period. Genotyping was performed using DNA extracted in a similar manner to iPSYCH2012 at Statens Serum Institut, Copenhagen DK using the Illumina Global Screening Array v2.0 (Illumina, San Diego CA, USA).

The trios dataset used in this study contains 128 parent-offspring trios where the offspring were ascertained for diagnoses of autism or attention-deficit/hyperactivity disorders with both parents born in Denmark, on or after May 1st 1981. These samples were genotyped using both the Infinium PsychChip v1.0 and the Illumina Global Screening Array v2.0.

Information on psychiatric diagnoses were obtained from the Danish national psychiatric central registers[31,32], demographic information including age, gender and parental birth place were obtained from the Danish civil registers[33,34].

The Personal Genomes Project - UK (PGP-UK) is an open source initiative aimed at facilitating access to multi-omics datasets for the purpose of gaining insights into biological and medical processes[29] and contains 1,100 citizens or permanent residents of the United Kingdom who passed a test aimed at educating them on the risks of sharing personal genetic data and further provided consent. DNA extraction was done from blood and whole genome sequencing performed using the Illumina HiSeq X Ten at an average depth of 15x. The resulting BAM files were deposited to the European Nucleotide Archive (ENA, study identifier: PRJEB17529).

Ethical Permissions

Research using iPSYCH and the trio data has been approved by the scientific Danish ethics committee, Danish health authority and the Danish neonatal screening biobank committee. PGP-UK has been approved by the University College London scientific ethics committee. All analysis was performed on a secure server within the Danish national life science supercomputing cluster (https://computerome.dtu.dk/).

Genotype Quality Control

Genotype data from iPSYCH2012, iPSYCH2015i, trios and PGP-UK were aligned to HRC v1.1 using genotype harmonizer version 1.4.20-SNAPSHOT[35]. SNPs not genotyped in all waves/batches of iPSYCH were excluded. Further filtration steps included exclusion of SNPs missing in at least 5% of the study subjects, SNPs showing differential missingness between cases and controls, SNPs failing tests of Hardy Weinberg equilibrium in controls of a homogenous genetic origin, SNPs significantly associated with a genotyping batch or wave, SNPs with minor allele frequencies less 0.001. A further 10,000 SNPs were masked to serve as a truth set for benchmarking the performance of imputation. Samples were excluded for abnormal levels of heterozygosity that cannot adequately be explained by admixture, missing greater than 5% of the SNPs that pass quality control. In case of duplicate samples, monozygotic twins, the sample with lower missingness was retained. This resulted in a total of 80,876 individuals genotyped at 251,551 SNPs in iPSYCH2012 and 48,974 individuals genotyped at 450,445

SNPs in iPSYCH2015i passing all quality control steps. The quality control steps are detailed in depth in supplementary sections S1-S3 and supplementary figure 1. All quality control steps were performed using PLINK v1.90b3o 64-bit 20 May 2015[36].

Pre-phasing Data Integration Protocols

We evaluated four different ways of integrating legacy data as shown in figure 1.

Separate. In this protocol as demonstrated in figure 1a, the samples from iPSYCH2012 and iPSYCH2015i are phased and imputed separately. The 124 trio offspring without the parental genotypes were added to both cohorts. Ten whole genome sequenced samples from the PGP-UK cohort were down-sampled to both the iPSYCH2012 and iPSYCH2015i SNPs that passed QC and merged with both cohorts. This resulted in two cohorts: (1) Cohort2012 (81,022 samples, 251,551 SNPs, 0.1% missingness) which includes iPSYCH2012, trio offspring genotyped on the Infinium PsychChip v1.0 and the whole genome sequenced PGP-UK samples down-sampled to the Infinium PsychChip v1.0 variants that pass QC. (2) Cohort2015i (49,120 samples, 450,455 SNPs, 0.31% missingness) which includes the iPSYCH2015i, trio offspring genotyped on the Illumina Global Screening Array v2.0 and the whole genome sequenced PGP-UK samples down-sampled to the Illumina global screening array v2.0 variants that pass QC.

Intersection. In this protocol as demonstrated in figure 1b, the samples from iPSYCH2012, iPSYCH2015i were merged together at the 116,962 SNP loci present on both the Infinium PsychChip v1.0 and the Illumina Global Screening Array v2.0 and passing quality control. 62 offspring samples were chosen at random from each of the trio datasets genotyped using both the Infinium PsychChip v1.0 and the Illumina global screening array v2.0 and merged to this dataset along with the 10 PGP-UK whole genome sequenced samples down-sampled to the 116,962 loci in common to both genotyping arrays. This resulted in the *intersection* (129,886 samples, 116,962 SNPs, 0.17% missingness) cohort.

Union. In this protocol as demonstrated in figure 1c, the samples from iPSYCH2012, iPSYCH2015i were merged with missingness to the 596,028 SNP loci genotyped on either the Infinium PsychChip v1.0 or the Illumina Global Screening Array v2.0 and passing quality control. To this, 62 samples from the trio dataset genotyped on both arrays were merged, same as in the INTERSECTION cohort. Five PGP-UK whole genome sequenced individuals were down-sampled to the SNPs present

on each genotyping array and merged resulting in the *union* cohort (129,886 samples, 596,028 SNPs, 44.54% missingness).

Twostage. In this protocol as demonstrated in figure 1d, the eight sets of phased haplotypes from the Cohort2012 and Cohort2015i obtained in the SEPARATE protocol were initially imputed using BEAGLE5.1 in batches of 10,000 samples to the 596,028 SNPs genotyped on either the Infinium PsychChipv1.0 or the Illumina Global Screening Array v2.0 and passing quality control with HRCv1.1 as the reference. Then the two cohorts were merged together, retaining the same 62 trio samples from each cohort as chosen in the INTERSECTION and UNION approaches and 5 PGP-UK samples from either cohort to form the *twostage* cohort (129,886 samples, 596,028 SNPs, 0% missingness).

All datasets were stored in variant call format (VCF) files (http://samtools.github.io/hts-specs/VCFv4.2.pdf) and manipulations were done using bcftools[37].

Phasing

Each cohort was phased using each of three tools, BEAGLE5, SHAPEIT4.1.2, EAGLE2.4.1, each at two different parameter values. The phasing using BEAGLE5 was performed using the default parameters and a higher resolution parameter, doubling the number of phase states to 560. Phasing using SHAPEIT4.1.2 was performed using the default parameter set and further at a higher resolution by doubling the pbwt-depth to 8. Similarly, phasing using EAGLE2.4.1 was performed using the default parameter set and further by doubling the parameter Kpbwt to 20000. The aim of this exercise is to benchmark the improvement in phasing accuracy at a higher resolution parameter set at the expense of longer run times and memory requirements. A CONSENSUS haplotype set was generated by taking the majority haplotype estimate across all three tools at both the default and higher resolution parameters at each locus in each individual using the consensusvcf module in the BEAGLE utilities tool set (consensusvcf.jar). The haplotype reference consortium HRCv1.1 dataset, consisting of 64,976 haplotypes[38].

Imputation

All cohorts were imputed using BEAGLE5.1 and HRCv1.1 as the reference. Due to the sample sizes, imputations were carried out in batches of 10,000 samples. Imputed dosage (DS) for an individual at a bi-allelic locus is calculated as DS = p(RA) + 2*p(AA) where p(RA) is the genotype

probability corresponding to the presence of one alternate allele as per the reference panel and p(AA) corresponds to the genotype probability of the presence of two copies of the alternate allele.

Evaluation of phasing accuracy

The accuracy of phasing was evaluated by calculating switch error rates at the heterozygous SNP loci in the trio offspring that are common to both iPSYCH genotyping arrays and pass quality control. A switch error occurs when there arises an inconsistency between the computationally assigned phase and the phase observed by mendelian transmission with knowledge of maternal and paternal haplotypes, suggesting that the alleles were switched between the two haplotypes in the offspring by the computational phasing. Switch error rate is calculated as the number of such switches divided by the total number of possibilities for switches[39]. The code for switch error rate calculation has previously been used in benchmarking phasing performance[9] and available on Github (https://github.com/odelaneau/switchError).

Evaluation of imputation accuracy

The imputation accuracy within the iPSYCH samples was evaluated as the squared Pearson correlation coefficient between the true genotypes and the imputed dosages within different minor allele frequency bins in HRCv1.1 at each of the 10,000 SNPs masked prior to phasing. The imputation accuracy within the UK personal genomes project samples was evaluated by calculating the squared pearson correlation coefficient between the true genotypes obtained from multi-sample variant calling using samtools[37] and the imputed dosage in different minor allele frequency bins in HRCv1.1 at 6,517,513 loci that were not genotyped on either iPSYCH array.

Imputation accuracy in non-European and admixed samples

To evaluate the variations in imputation accuracy by ancestral origin, the iPSYCH samples were grouped according to the country of birth of both parents as obtained from the Danish civil registers[33,34]. Within each ancestry group, imputation accuracy was calculated as the squared Pearson correlation coefficient between true genotypes and imputed dosages at different minor allele frequency bins in HRCv1.1.

Evaluation of directional bias in imputation

To evaluate the potential of direction biases in estimates of imputed dosages, the bias at each marker was calculated as the difference between the true minor allele count or a particular sample and the estimated dosage at each of the 10,000 masked SNPs. The means of these biases were calculated within different minor allele frequency bins.

Polygenic Scores

Polygenic scores (*PGS*) for each individual, *j*, were constructed using simulated per-allele effects as follows:

$$PGS_j = \sum_{i=1}^{m} \beta_i X_{ij}$$

where *m* is the total number of SNPs (10,000 masked SNPs), $\boldsymbol{\beta}_i$ is simulated effect for SNP *i*, X_{ij} is the imputed dosage, best guess, or genotyped count of effect alleles for individual j at SNP i. The effect sizes ($\boldsymbol{\beta}_i$) were generated by simulating a continuous phenotype with a SNP heritability of 0.5 from the 10,000 masked SNPs as causal loci using the GCTA[40]. Variance explained in the PGS was calculated by fitting a linear model with the phenotype as the outcome and the PGS per individual calculated using the true genotypes and imputed dosages as explanatory variables.

Association Tests

To evaluate the presence of batch artifacts in each protocol we conducted multiple GWAS using the glm module of PLINKv2.00a2LM 64-bit Intel (10 Nov. 2019)[41]. The tests were performed in multiple versions of the combined iPSYCH data, with the iPSYCH cohort as the outcome. As a baseline truth we performed the GWAS for cohort membership in the combined data set of the true values of 10,000 masked genotypes. The first set of test data compared each set of imputed genotypes in one cohort to true genotypes in the other cohort. The second set of test data compared imputed genotypes of each cohort (i.e., iPSYCH2012 vs. iPSYCH2015i) within each protocol (e.g., *separate*, *union*, *intersection*, *twostage*), where the data, with the exception of the separate protocol, had been merged prior to phasing. Tests were restricted to individuals without mental disorders (i.e., a random sample of psychiatric controls), of a homogenous genetic origin based on principal component analysis (Supplementary S1) using Eigenstrat[42], and pruned for relatedness beyond the third degree using

kinship coefficients estimated by KING[43]. The true genotypes or imputed dosages were used as explanatory variables and genotyping array (i.e., iPSYCH2012 vs. iPSYCH2015i) as the outcome. The overall inflation of test statistics above the null was evaluated using the genomic inflation factor which compares the median of the chi-square test statistic obtained from each GWAS to the expected median of a chi-square distribution with 1 degree of freedom.

Results

Phasing Accuracy:

Phasing accuracy was measured as switch error rates (SER; Methods) comparing ground truth offspring haplotypes phased using parental information to those estimated computationally by EAGLE2.4.1, SHAPEIT4.1.2, and BEAGLE5.1, at two parameter settings, and across the four data integration protocols (Figure 2a, Supplementary S3, Supplementary Table 10). The accuracy of phasing depends on the integration protocol and the phasing tool along with its associated parameters, although to different extents. The performances appear related to differences in target sample size, SNP density in the target dataset, and rate and structure of initial missing genotypes.

In general, the *twostage* protocol showed the lowest SER across phasing tools, while the intersection protocol showed the highest SER. The ranking of the protocols was generally consistent across tools, with the exception of the *union*, which achieved the lowest overall SER with BEALGE5 at 560 phase states. Surprisingly, the *union* was also the worst performing protocol when taking consensus haplotypes, suggesting the structured missing data introduced by this protocol in the initial genotypes can cause systematic phasing errors across tools.

Broadly, BEAGLE5 and SHAPEIT4.1.2 performed similarly in terms of SER, both outperforming EAGLE2.4.1 methods and parameters such that the worst phasing by these tools (*intersection*) was comparable to the best from EAGLE2.4.1 (*twostage*). The *union* was again a point of departure from the trends with BEAGLE5 performing better on the *union* and SHAPEIT4.1.2 performing better on the *twostage*. This again implicates issues around structured missing data in the initial genotypes in phasing performance, and suggests BEAGLE5 has a more robust handling than does SHAPEIT4.1.2. When considering the *twostage* protocol, which we hypothesized could mitigate initial missing genotypes, SHAPEIT4.1.2 performance was comparable to BEAGLE5's performance on the *union* (and

better than on the *twostage*), suggesting, at least at these tested parameters and modulo initial missing data, SHAPEIT4.1.2 may have a slightly better performing phasing algorithm.

These differences in performance for phasing can be related to variability in the density of SNPs, numbers of subjects, and amount of missing data across integration protocols. Comparing the phasing accuracy across chromosomes within each method and data integration protocol reveals that the accuracy of phasing follows the number of SNPs per centimorgan in the target dataset, with denser chromosomes across protocols and protocols resulting in denser SNP sets across chromosomes showing lower SER (Supplementary figure 1). We also observe that certain tools perform better with larger target sample sizes, provided the SNP density in the target dataset is not too sparse, whereas others perform better with higher SNP density, as observed in comparison between the two cohorts of the *separate* protocol. EAGLE2.4.1 and BEAGLE5 produce more accurate estimates in the Cohort2012 where the sample size is higher and SNP density is sparser whereas SHAPEIT4.1.2 produces more accurate estimates in the Cohort2015i where the SNP density is higher while the target sample size is comparatively smaller. As mentioned above, the worse performance of SHAPEIT4.1.2 and EAGLE2.4.1 on the *union* as opposed to the *twostage* highlight the sensitivity to initial missing data.

Imputation Accuracy:

The accuracy of imputations derived from each set of haplotype scaffolds (i.e., from each tool and data integration protocol pair) are presented in figures 2b,c and Supplementary Table 7. Most of the variability in imputation performance stems from the choice of data integration protocol. When measuring the empirical correlation between the true masked genotype and the imputed dosage, the best imputation is obtained when the cohorts are phased separately. This trend is consistent across scaffolds generated by all tools. The added bioinformatics effort aimed at enhancing sample size without missingness with the *twostage* protocol does not yield a higher imputation accuracy than the *separate* protocol. The imputation accuracy is degraded when using the *intersection* protocol for data integration with an attenuation between 8.4-13.6% at common allele frequencies and 13.9-18.6% at rare minor allele frequencies when compared to the *separate* protocol. Again, haplotypes estimated by SHAPEIT4.1.2 in the *union* protocol are an outlier and resulting imputations are of noticeably poorer quality compared to haplotype estimates obtained by other tools. Phasing in the presence of missingness is a two step process where each phasing tool makes a rough imputation of missing data prior to haplotype estimation. If this data is carried through, and not overwritten in the imputation

step, the prephasing imputation algorithm implemented by SHAPEIT4.1.2 could be the reason for this performance issue. This hypothesis becomes more credible when considering the imputation accuracies obtained from the *twostage* protocol using SHAPEIT4.1.2, where the attenuation is mitigated and the pattern of results described above is replicated by the imputation of the PGP-UK samples (supplementary figure 3).

A comparison of imputation accuracies between Cohort2012 and Cohort2015i within the *separate* protocol using the PGP-UK samples (Supplementary Figure 2c, Supplementary Table 9) shows higher imputation accuracy in Cohort2015i with a higher SNP density as compared to Cohort2012 with a larger sample size with a difference as high as 6.7% in the minor allele frequency bin $MAF \in (0.1, 0.05]$. There does not appear to be much increment in imputation accuracy from haplotypes obtained using the enhanced parameters for higher haplotype accuracy as compared to the default parameters (Supplementary figures 2a,b). Taken together, these results show that the accuracy of imputation suffers when merging cohorts genotyped on different arrays prior to phasing and the choice of tool is much less relevant than the protocol for integrating data.

Due to the robustness of haplotypes estimated by BEAGLE5 across all data integration scenarios, we proceed with the imputed dosages generated using these haplotypes for all further analyses.

Evaluation of directional bias in imputation:

Most imputation accuracy assessments, as above, focus on the estimated or empirical correlations between true and imputed genotypes[44,45], which describe a reduction in genotypic variance and subsequent power. They do not assess or describe any potential directional bias in imputed genotypes, as was described in animal breeding examples[26]. We estimate the average difference in the imputed count minor alleles for each true genotype group (i.e., 0, 1, or 2 true copies) at the 10,000 masked SNPs (Table 1). Individuals carrying minor alleles have them under counted, at times quite severely. These differences persist after applying common quality filtering criteria typical in complex trait analyses.

Imputation accuracy in non-European and admixed samples

It is known that GWAS results and subsequent polygenic risk predictions constructed using them do not transfer well across populations[46]. Much of the focus on this debate has centered around inaccuracies in the estimation of the effect sizes due to variable linkage disequilibrium and minor allele frequencies across population[47], but non-European haplotypes are underrepresented in reference data sets and so imputed genotypes may also be of lower quality. To evaluate the accuracy of genotype imputation in non-European and admixed samples within iPSYCH, we compared the correlation between true and impute genotype dosage in subsets of individuals grouped by the birthplace of both parents (Figure 3a,b; Supplementary Figure 4, Supplementary Table 8). Individuals born to non-Scandinavian European parents had lower imputation accuracy (7.07-12.58%) than those with both parents born in Denmark. These effects were larger for individuals with both parents born in Asia (11.1-11.2%), Africa (17.37-17.48%), or the Middle East (11.2-17.7%). The average accuracy of genotype imputation in admixed individuals was intermediate, varying between 4.47-8.56%.

Reliability of imputation quality metrics:

The imputation software BEAGLE5.1 provides an estimated quality score for imputed dosages (BEAGLE r^2) at each SNP, which is a predicted correlation between the true and estimated genotypes at a given variant. The r^2 at an imputed locus is an important quantity because it can be used to estimate the reduction in effective sample size for an association test[48]. This value is also used as a filtering threshold to ensure only high quality markers are used for association tests and polygenic scoring. As such, we sought to compare the accuracy of this metric across data integration protocols by comparing it to the empirical imputation accuracy calculated for the 10,000 masked SNPs. The results suggest that Beagle's r^2 is best calibrated for *intersection* ($r^2_{\text{BEAGLE-r2, EMPIRICAL-r2}} = 0.98$) protocol but it tends to overestimate the accuracy of imputation in the *union* ($r^2_{\text{BEAGLE-r2, EMPIRICAL-r2}} = 0.77$) protocol. The Beagle and Empirical r^2 estimates of imputation quality appear to be broadly in agreement in the *separate* ($r^2_{\text{BEAGLE-r2, EMPIRICAL-r2}} = 0.9$) and *twostage* protocols ($r^2_{\text{BEAGLE-r2, EMPIRICAL-r2}} = 0.9$). (Supplementary Figure 3). Thus, the implications of errors in imputation quality metrics should be considered carefully when haplotypes scaffolds are generated from SNP sets containing a lot of missing data.

Impact on polygenic scores:

Polygenic scores sum the products of an estimated SNP effect, usually taken from a GWAS, with the effect allele count across a large collection of SNPs within an individual. Here we assessed the effects of loss of information and directional bias in imputed genotypes on simulated polygenic scores (Methods, Figure 4a, Supplementary S7). PGS calculated using true genotypes explain 44.93% of the variance in a simulated phenotype with 50% heritability and an attenuation in variance explained is observed when PGS were calculated using imputed dosages or best guess genotypes. To estimate the attenuation that can be attributed to imputation biases towards the major allele in the reference, the PGS was further decomposed into PGS when the effect allele is the major allele in HRCv1.1 and PGS when the effect allele is the minor allele. The percentage attenuation due to imputation as compared to using true genotypes is shown in figure 4b.

The attenuation in variance explained in the phenotype using PGS calculated with imputed dosages as compared to the true genotypes was lowest when using the *twostage* protocol at 5.21% and highest when using the *intersection* protocol at 13.22%. There appears to be a minor advantage to using imputed dosages as compared to best guess genotypes for PGS analysis as an attenuation of 6.54% was observed in PGS calculated using best guess genotypes in the *twostage* protocol and 16.2% in the *intersection* protocol as compared to PGS calculated using true genotypes. PGS decomposed on the basis of the allele frequency in the reference shows a 1% increment in attenuation when the effect is the minor allele in HRCv1.1 as compared to when it is the major allele. This difference is 1.14% when PGS is calculated using imputed dosages from the *intersection* protocol.

Another application of PGS is to calibrate the risk of an individual towards a trait by checking if they fall in the actionable top 5 percentiles of a risk distribution computed in an independent population. To evaluate the performance of imputations for such an application, PGS calculated using true genotypes and imputed dosages from the four data integration protocols for every control individual in iPSYCH2015 were used to place them into the percentiles they fall in the distribution of PGS calculated in iPSYCH2012 controls using true genotypes.

The results show 10.28% of the iPSYCH2015i individuals who are classified in the top decile of PGS in the iPSYCH2012 population using PGS calculated from true genotypes fall out of this range when the PGS is calculated using imputed dosages from the *separate* protocol. This drop out reaches 10.4% when the PGS is estimated with imputed dosages arising from the t*wostage* protocol, 13.88%

when using imputed dosages from the *union* protocol and 22.41% when choosing imputed dosages from the *intersection* protocol. These results replicate findings from animal breeding studies[26] demonstrating the bias introduced by imputations in genetic predictions.

Batch effects:

Batch artifacts associated with component genotyping arrays may arise when combining data imputed by different protocols. Association studies were performed comparing genotypes and imputed dosages at the masked SNPs from all four data integration protocols in unrelated controls of iPSYCH2012 and iPSYCH2015i of a homogenous genetic origin with the genotyping array as the outcome (see Methods). The resulting genomic inflation factor in test statistics across different thresholds for imputation quality is shown in figure 5a, supplementary S8. The number of SNPs used in the association tests at each threshold for imputation quality is shown in figure 5b.

The baseline for the inflation observed by comparing the genotyped SNPs in controls is λ_{gc} = 1.05. There appears to be no inflation from the *intersection* protocol when SNPs are imputed in both iPSYCH2012 and iPSYCH2015i. The inflation in test statistics is highest when using the *union* protocol. Using the *separate* and *twostage* protocols, the inflation is reduced at high thresholds for imputation r^2 obtained from BEAGLE, but is not completely eliminated. For example, the λ_{gc} at r^2 >= 0.9 for the *separate* protocol is 1.13 when comparing SNPs genotyped in iPSYCH 2012 to SNPs imputed in iPSYCH 2015, 1.18 when comparing SNPs imputed in iPSYCH2012 to SNPs genotyped in iPSYCH2015i and 1.1 when comparing SNPs imputed in both. At this threshold approximately 22% of the imputed SNPs are excluded. This analysis suggests that imputations performed from different genotyped backbones, which result in genotyped SNPs being compared to imputed SNPs, will contain batch artifacts that can be difficult to remove by standard SNP exclusion.

Discussion

As the cost of genotyping drops, the burden of complex trait analysis has moved away from data generation and towards storage, computational requirements and the requisite bioinformatics expertise to integrate and analyze large datasets[49]. Developing bioinformatics protocols in practice is about informed trade-offs and not single feature optimization, which is not realistic, especially when considering expensive legacy data. The comparisons presented in this study were intended to benchmark protocols in real data integration scenarios, where often, it is not possible to optimize any

one aspect of the data (e.g., SNP density, sample size, missingness) without making trade-offs on others.

Phasing and imputation have remained a sort of black box in bioinformatics pipelines with researchers having the opportunity to avail themselves of services like the Michigan imputation server[50] to reduce the computational burden of data preparation for complex trait analysis. However, the privacy stipulations governing datasets generated through national biobanks might prohibit the use of such services. The benchmarking work presented in this study stresses the importance of making an educated choice of phasing tool when choosing data integration protocols, that could introduce a tradeoff among peculiarities such as a sparse marker set, small sample size, high missingness in the input dataset, or the potential of batch artifacts.

The benchmarking of imputation accuracy presented in this study replicates previous findings[22], suggesting imputation from the intersection of markers when incorporating samples genotyped on multiple arrays leads to a loss of accuracy while imputation from the union of the markers leads to spurious associations with genotyping arrays[22]. The haplotype reference panels that are currently employed for phasing and imputation are skewed towards Europeans and the evaluation of imputation within study individuals grouped by parental birthplace shows differentially worse accuracy in non-Europeans, stressing the need for reference panels with a more genetically diverse catalog of haplotypes, if genotyping arrays and imputation are to be used in precision medicine initiatives in a fair and equitable manner[46,47]. While our comparisons held the reference population constant to the largest set of haplotypes that are currently publicly available, testing the imputation performance with varying references would also be informative. There has been demonstrated improvement in imputation quality for individuals of Hispanic/Latin and African descent using the NHLBI Trans-Omics for Precision Medicine whole genome sequenced reference panel[51]. However, this reference panel is currently only available through an imputation server, rendering its usage prohibitive for studies with data privacy stipulations.

Imputed data will contain non-random errors, especially when there is systemic missingness in the data being imputed, as can be the case when genotyping of samples is performed in batches, using biological data collected over different time periods and therefore it is critical to consider the sensitivity of any analysis performed on these datasets. Technical artefacts in the genotype generation process is one of the sources of poor performance of polygenic scores across cohorts[23]. The demonstrated bias of imputed dosages towards the major allele in the haplotype reference as well as

the attenuation introduced when polygenic scores are calculated using imputed dosages and further best guess genotypes as compared to genotyped SNPs is under-researched in studies of human populations. The benchmarking done in this study stresses the need to account for the proportion of imputed to genotyped SNPs in cohorts when comparing the PGS of individuals across cohorts genotyped using different arrays. One of the clinically informative ways to use polygenic scores is to select the subset of individuals in the extreme deciles of a PGS distribution to be prioritized for clinical intervention[23,27]. The discordance between individuals in the tails of the PGS distribution when the PGS is calculated using true genotypes as compared to imputed dosages demonstrates the impact errors introduced during phasing and imputation could have on genetic risk profiling in clinical settings.

The presence of false positive associations with genotyping arrays as observed by the inflation in test statistics when comparing allele frequencies of genotypes and imputed dosages between cohorts genotyped using different arrays as demonstrated in this study shows the need to pick stringent quality control thresholds to avoid a lack of replication or possible false positives in genome wide association studies. While the stringent filtering might reduce the power to detect association due to exclusion of a large number of imputed markers, other approaches such as accounting for the biases towards genotyping array by including the genotyping array as a covariate in regression models or as a fixed effect in linear mixed models need to be further investigated.

In conclusion, this study demonstrates four different ways of integrating legacy data genotyped on multiple arrays. Empirical evaluations of the impact of each data integration protocol over the accuracy of haplotype estimation, missing data imputation and the analysis of complex traits suggests that the best protocol is to phase and impute the data separately when the genotyping arrays used do not have a good overlap of markers.

Table 1. Imputed data shows a directional bias towards the major allele. Showing the average imputed dosage corresponding to true count of minor allele at different minor allele frequency bins across the four data integration protocols.

PROTOCOL	Truth	Estimated number of minor alleles					
		MAF <= 0.01		0.01 < MAF <= 0.1		MAF > 0.1	
		ALL	DK; r^2 >= 0.8	ALL	DK; r^2 >= 0.8	ALL	DK; r^2 >= 0.8
SEPARATE	0	0.0013	0.0007	0.0053	0.0026	0.0237	0.0161
	1	0.6963	0.8568	0.9314	0.9679	0.9831	0.9891
	2	1.2719	1.5554	1.8547	1.9299	1.939	1.96
INTERSECTION	0	0.0019	0.0007	0.0135	0.0048	0.0697	0.0327
	1	0.582	0.8592	0.8391	0.9491	0.9528	0.978
	2	1.0408	1.5554	1.649	1.889	1.8323	1.9208
UNION	0	0.0014	0.0009	0.0057	0.0042	0.0272	0.022
	1	0.6672	0.8019	0.9152	0.9366	0.9794	0.983
	2	1.192	1.473	1.818	1.8718	1.9304	1.9436
TWOSTAGE	0	0.0013	0.0007	0.0053	0.0026	0.0246	0.0161
	1	0.6963	0.8565	0.9317	0.9679	0.9831	0.9891
	2	1.2714	1.5577	1.8551	1.9299	1.9392	1.96

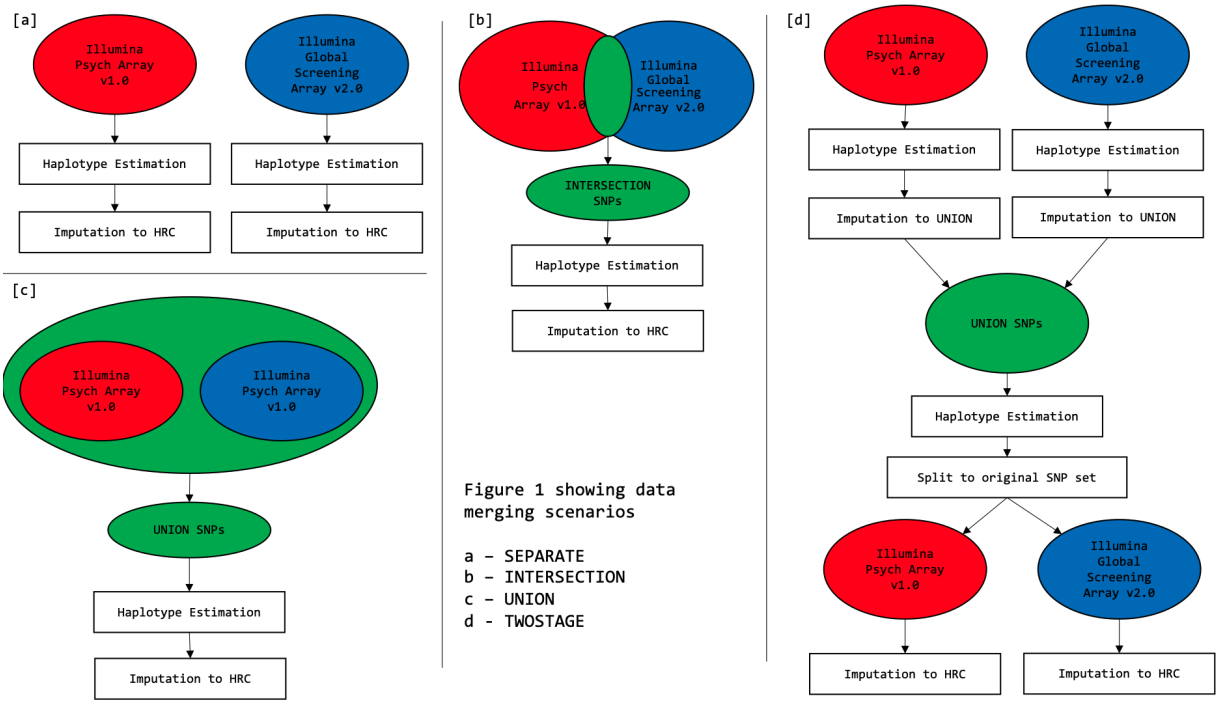


Figure 1. Four pre-phasing data integration protocols. [a] shows the *separate* protocol where the cohorts genotyped on each array are phased and imputed separately. [b] shows the *intersection* protocol where the two cohorts are merged to include only SNPs in common to both genotyping arrays prior to phasing and imputation. [c] shows the UNION protocol where the two cohorts are merged to include SNPs genotyped on either array and the resulting dataset with missingness is phased and imputed. [d] Shows the two stage protocol where the haplotypes obtained from the *separate* protocol are initially imputed to the markers in the *union* protocol, prior to a second stage of phasing before the cohorts are split back to the original sets of genotyped SNPs after which imputation to the full reference panel is performed.

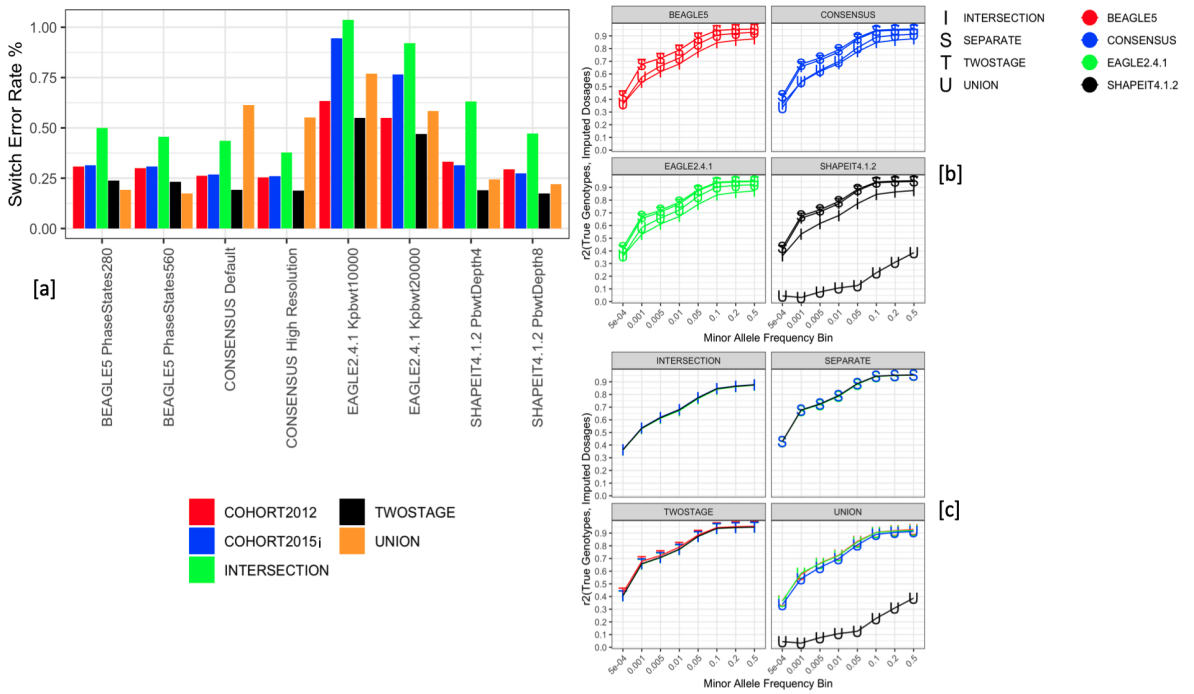


Figure 2. Phasing and imputation accuracy vary across state of the field tools and data integration approach. [a] Shows the accuracy of phasing across the three tools at two parameter sets each and a *consensus* approach taking the majority haplotype at each locus from the three tools at both parameter sets across all four data integration protocols. Default parameters are SHAPEIT4.1.2 pbwt-depth=4, BEAGLE5 phase-states=280, EAGLE2.4.1 Kpbwt=10000. High Resolution parameters are SHAPEIT4.1.2 pbwt-depth=8, BEAGLE5 phase-states=560, EAGLE2.4.1 Kpbwt=20000. The switch error rates were computed within 124 trio offspring by comparing the computationally assigned phase to the mendelian transmission from known parental genotypes at the heterozygous loci common to both genotyping arrays.[b] Shows the imputation accuracy within each data integration protocol, grouped by choice of phasing tool at different minor allele frequency bins at the 10,000 SNPs common to both genotyping arrays that were masked prior to phasing. [c] shows the accuracy of imputation from haplotypes estimated using the three different tools and the consensus approach grouped by choice of data integration protocol at different minor allele frequency bins. All imputations were performed using BEAGLE5.1 with the HRCv1.1 as the haplotype reference panel.

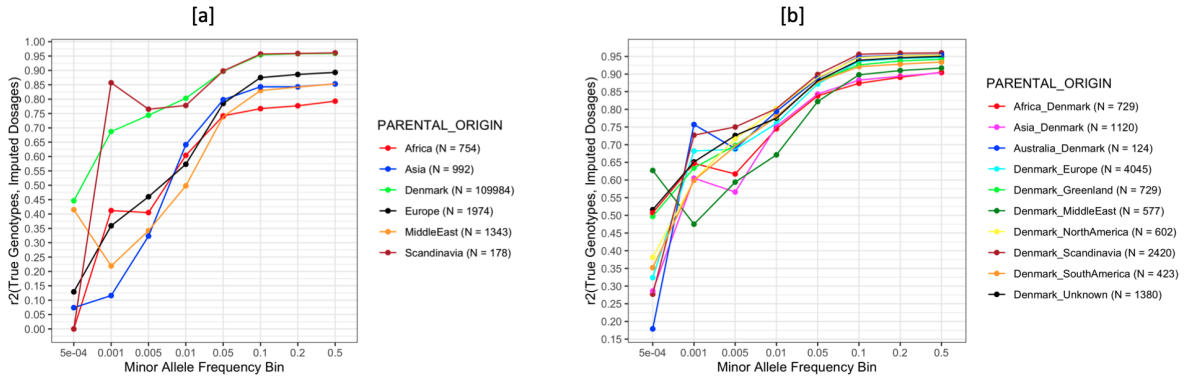


Figure 3. The accuracy of imputation varies extensively by genetic ancestry. [a] shows the imputation accuracy in iPSYCH samples grouped by parental birth place as ascertained from the Danish civil registers at different minor allele frequency bins within the 10,000 SNPs common to both genotyping arrays, masked prior to phasing. [b] shows the imputation accuracy in admixed samples where at least one parent was born in Denmark. All imputations were performed using the *separate* protocol, haplotype estimation was performed using BEAGLE5 phase-states=560, imputations were performed using BEAGLE5.1 with the HRCv1.1 as the reference.

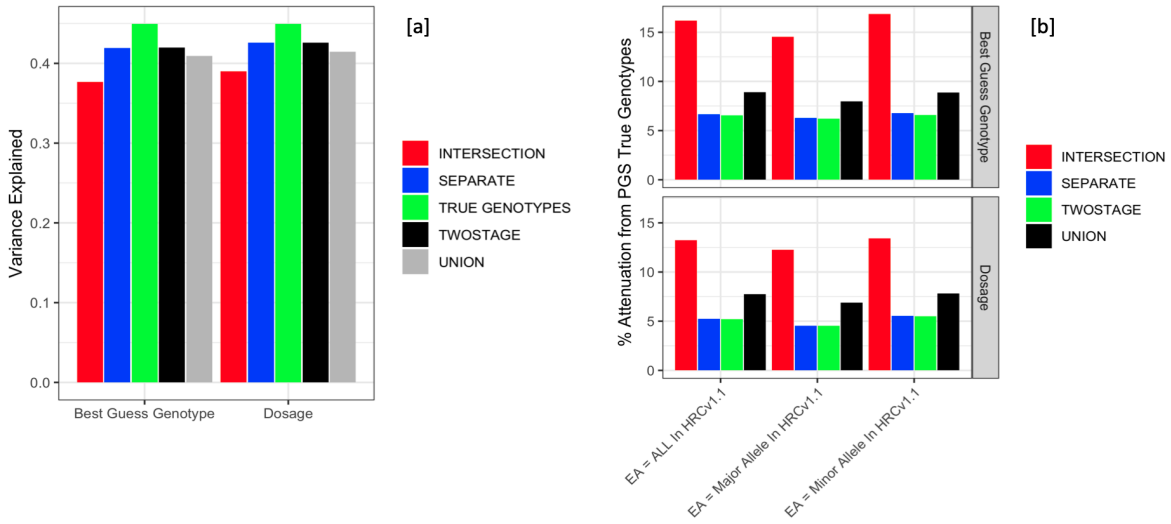


Figure 4. The effects of polygenic scores are attenuated when using imputed data. [a] shows the variance explained in a simulated continuous phenotype with a SNP heritability of 0.5 and the 10,000 SNPs common to both genotyping arrays, masked prior to phasing as causal loci. Variance explained was calculated using the true genotypes, along with imputed dosages and best guess genotypes from the four different data integration protocols. Haplotypes were phased using BEAGLE5 phase-states=560, imputations were done using BEAGLE5.1 with the HRCv1.1 as the reference. [b] Shows the percentage attenuation in variance explained by the PGS calculated using imputed dosages and best guess genotypes as compared to PGS calculated using true genotypes across four different data integration protocols. The attenuation is further decomposed into the attenuation arising when the effect allele is the major allele or minor allele in the HRCv1.1 reference panel.

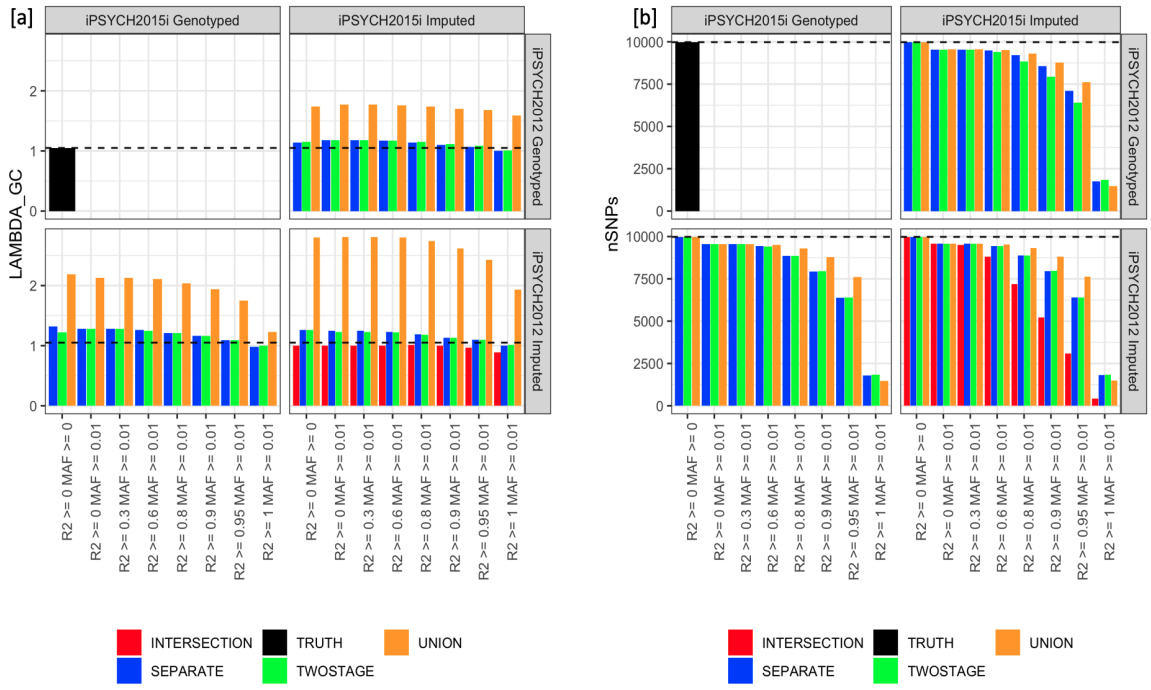


Figure 5. Inflation of test-statistics shows type-I errors associated with imputation. [a] Shows the inflation in test statistics represented using lambda genomic control, when performing an association test at each of the 10,000 SNPs common to both genotyping arrays masked prior to phasing. Controls of a homogeneous genetic origin were compared between the iPSYCH2012 and iPSYCH2015i cohorts with the genotyping array as the outcome at different thresholds of post-imputation quality control across the four different data integration protocols. The dotted horizontal line indicates the baseline λ_{gc} when the association test was performed using true genotypes from both arrays. Haplotypes were phased using BEAGLE5 phase-states=560, imputations were done using BEAGLE5.1 with the HRCv1.1 as the reference. [b] Shows the number of SNPs left after each threshold of post imputation quality control across the four data integration protocols.

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
4. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
5. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
6. Zeggini, E. & Ioannidis, J. P. A. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191–201 (2009).
7. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).
8. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
9. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
10. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
11. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
12. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* vol. 3 160025 (2016).
15. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285–1295 (2015).
16. Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N. & Ioannidis, J. P. A. The power of meta-analysis

in genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **14**, 441–465 (2013).

17. Zaitlen, N. & Eskin, E. Imputation aware meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **34**, 537–542 (2010).
18. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
19. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
20. Sinnott, J. A. & Kraft, P. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum. Genet.* **131**, 111–119 (2012).
21. Uh, H.-W. *et al.* How to deal with the early GWAS data when imputing and combining different arrays is necessary. *Eur. J. Hum. Genet.* **20**, 572–576 (2012).
22. Johnson, E. O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum. Genet.* **132**, 509–522 (2013).
23. Ni, G. *et al.* A comprehensive evaluation of polygenic score methods across cohorts in psychiatric disorders. *Genetic and Genomic Medicine* (2020) doi:10.1101/2020.09.10.20192310.
24. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
25. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
26. Pimentel, E. C. G., Edel, C., Emmerling, R. & Götz, K.-U. How imputation errors bias genomic predictions. *J. Dairy Sci.* **98**, 4131–4138 (2015).
27. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
28. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
29. Chervova, O. *et al.* The Personal Genome Project-UK, an open access resource of human multi-omics data. *Sci Data* **6**, 257 (2019).
30. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
31. Munk-Jørgensen, P. & Mortensen, P. B. The Danish Psychiatric Central Register. *Dan. Med. Bull.* **44**, 82–84 (1997).
32. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
33. Pedersen, C. B. The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).

34. Schmidt, M., Pedersen, L. & Sørensen, H. T. The Danish Civil Registration System as a tool in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).
35. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
36. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
37. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
38. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
39. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).
40. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
41. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
42. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
43. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
44. Shi, S. *et al.* Comprehensive assessment of genotype imputation performance. *Human* (2018).
45. Schurz, H. *et al.* Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Front. Genet.* **10**, 34 (2019).
46. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
47. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
48. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
49. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
50. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

51. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).

SUPPLEMENTARY INFORMATION

S1. QUALITY CONTROL OF GENOTYPING DATA

The quality control steps prior to phasing are divided into two stages. An initial SNP level QC and a second sample level QC performed on a subset of individuals of a homogenous genetic origin within the iPSYCH sample.

S1.1. Identifying a genetically homogenous sample subset for QC:

Certain steps in the quality control process such as tests of Hardy Weinberg equilibrium, identification of samples with abnormal heterozygosity etc could be biased by genetic diversity in the dataset. In order to perform these quality control steps in an unbiased manner, we identify a set of samples of a homogenous genetic origin. In order to do this, the variant calls from the 1000 genomes phase 3 project[1] were downloaded in VCF format.

Within each sub-population of the 1000 genomes dataset, we excluded variants for the following reasons:

- Less than 5% minor allele frequency
- Hardy Weinberg $p < 10\text{-}6$
- Pairwise $r2 > 0.1$ in a 1kB region
- No overlap with the marker set in the Infinium Psych Chip v1.0 and the Global Screening array v2.0.
- Insertions/Deletions
- Regions with extended linkage disequilibrium[2].

The resulting data was merged with iPSYCH2012 and iPSYCH2015i using PLINK[3]. We performed a principal component analysis using the smartpca module of the eigensoft software package[4], the principal components were computed using the 1000 genomes samples and the iPSYCH2012, iPSYCH2015i samples were projected into the resulting principal component space.

We further utilized the Danish national birth records to identify a set of 47,586 individuals whose parents and both sets of grandparents were born in Denmark. For each sample in our dataset, we

calculate the Mahalanobis distance of the sample from the multivariate mean of the joint distribution of the first ten principal components obtained from the 47,586 individuals previously identified. We exclude a sample as an outlier if the distance has a probability less than 5.73×10^{-7} under a chi-square distribution with 10 degrees of freedom. This resulted in 120890 samples classified as inliers to be used for quality control.

S1.2. SNP QC

S1.2.1 Aligning to the reference:

All 26 waves of iPSYCH2012, 78 waves of iPSYCH2015i, Trios2012, Trios2015 and the PGP-UK samples were aligned to Haplotype Reference Consortium v1.1 (hereafter referred to as HRC) using GenotypeHarmonizer v1.4.20-SNAPSHOT[5]. SNP IDs in the target datasets were harmonized to the SNP IDs in the HRC where a match was found, A/T and G/C SNPs were rescued where possible using linkage disequilibrium information, variants not present in the reference, multi-allelic SNPs and indels were excluded.

S1.2.2 SNP Missingness:

Per SNP and sample missingness were calculated using PLINK 2.0. Genotyping for iPSYCH2012 was performed in 26 waves. We initially excluded variants missing in > 5% of samples in each individual wave. The samples were further merged and variants that were either not genotyped in all 26 waves or were found to be missing in >= 5% of samples in the merged dataset were further excluded. 344,498 SNPs pass this QC.

The genotyping for iPSYCH2015i was performed in 78 waves. We excluded SNPs missing in excess of 5% of the samples in each genotyping wave. Samples were merged across batches and SNPs missing in more than 5% of samples across the entire cohort were removed. A total of 558,013 SNPs pass missingness filters.

S1.2.3 Differential Missingness between Cases and Controls:

We test for SNPs showing differential missingness between cases and controls of a homogenous genetic origin as described in section 1 using the --test-missing option in PLINK. We excluded SNPs

that show evidence for differential missingness with an FDR adjusted p-value <= 0.2. 342,837 SNPs in iPSYCH2012 and 555,131 SNPs in iPSYCH2015i pass this filter.

S1.2.4 Test of Hardy Weinberg Equilibrium in Controls:

The individuals of a homogenous genetic origin as derived in section 1 were further subset to include individuals without any disease diagnosis as ascertained from the Danish national patient registers and a test for Hardy Weinberg equilibrium was performed using the --hardy option in PLINK. We exclude SNPs that fail this test with an FDR adjusted p <= 0.2. 338,104 SNPs in iPSYCH2012 and 544,308 SNPs in iPSYCH2015i pass this QC.

S1.2.5 SNPs significantly associated with a genotyping wave or batch:

Due to the large sample size of iPSYCH, the genotyping for iPSYCH2012 was performed in 26 waves and the genotyping for iPSYCH2015i was performed in 78 waves. To identify markers showing significant batch effects, we performed 26 and 78 logistic regressions in iPSYCH2012 and iPSYCH2015i respectively where samples of a homogenous genetic origin in a particular wave are cases and samples in other waves are controls. For each SNP, we take the minimum of p-values from all association tests.

The p-values thus selected do not follow a uniform distribution and the cumulative distribution function of drawing minimums from n independent distributions $Y = \min(p_1, p_2, .. p_n)$ is given by

$$CDF(Y) = p(Y <= y) = 1 - 1(1 - y)^n$$

If p_i is the i^{th} element in a set of m sorted p-values, the CDF of p_i is given by i/m. The i^{th} element in a set of m sorted minimum p-values is given by

$$p_i = 1- (1 – 1/m)^{1/n}$$

The qq-plot of observed vs expected p-values using the above theoretical distribution suggests some inflation.

FDR adjustment using the above CDF is given by

$$p_{fdr} = m - (1 - p_i)^n / sum(p < p_i)$$

We chose an FDR adjusted p-value cut-off of 0.1 to exclude SNPs, which corresponded to a p-value of 6.31×10^{-5} in iPSYCH2012 and 2.38×10^{-6} in iPSYCH 2016. SNPs passing QC filters, iPSYCH 2012 = 333,308, iPSYCH2015i = 543,422.

S1.2.6 Minor Allele Frequency:

A subset of 34,545 individuals in iPSYCH2012 were exome sequenced using the Illumina capture kit on HiSeq machines. Quality control was performed using HAIL and variant calling was performed in accordance with the GATK best practices. More details on the data processing is described elsewhere[6].

For these individuals, we calculated genotype concordance between the exome sequencing data and genotypes from the iPSYCH2012 array data using bcftools[7] as shown in supplementary table 1. We observe that the concordance between genotyped and next generation sequencing datasets drops sharply at minor allele frequencies below 0.001. So we chose this as a sensible threshold for censoring SNPs. SNPs passing QC filters: iPSYCH2012: 261,551, iPSYCH2015i: 460,445.

Allele Frequency Bin	**Concordance between genotyping array and Exome Sequencing Data**	**Number of SNPs**
0.00001 - 0.0001	0.4085	20701
0.0001 - 0.001	0.7976	30367
0.001 - 0.01	0.9676	14145
0.01 - 0.1	0.9966	6795
0.1 - 0.5	0.999	5081
0.5 - 1	0.9991	28

Supplementary Table 1. Concordance between genotypes from Infinium Psych Chip v1.0 and whole exome sequencing data in a subset of 34,545 individuals in iPSYCH2012.

S1.2.7 SNP Masking:

In order to evaluate the performance of missing data imputation, we randomly selected 10,000 SNPs that were genotyped on both the Illumina PsychArray v1.0 and the Illumina Global Screening Array v2.0 using the sample function in R. These were excluded prior to haplotype estimation. SNPs used for haplotype estimation and imputation, iPSYCH2012: 251,551, iPSYCH2015i: 450,445.

S1.3. SAMPLE QC

S1.3.1 Abnormal Heterozygosity:

Abnormal levels of heterozygosity that cannot adequately be explained by admixture, population structure or runs of homozygosity could indicate sample contamination. To identify individuals with heterozygosity that cannot be accounted for by population phenomena, we use an approach described by the UK biobank (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf). Per sample heterozygosity, homozygosity and missingness were calculated using PLINK --het, --homozyg and --missing options respectively. Ancestry adjusted heterozygosity is computed as the residuals from the model shown below:

$$H(x) \sim H_0 + PC_1 + PC_2 + PC_3 + PC_4 + PC_1^2 + PC_2^2 + PC_3^2 + PC_4^2 + PC_1*PC_2 + PC_2*PC_3 + PC_3*PC_4 + PC_4*PC_1 + PC_1*PC_3 + PC_2*PC_4 + E$$

Where H(x) = Observed heterozygosity
H_0 = Mean heterozygosity/Intercept
PC_1, PC_2, PC_3, PC_4 = First four principal components of genetic ancestry
E = Residual/Ancestry adjusted heterozygosity

We further fit two linear models predicting the observed and ancestry adjusted heterozygosities from runs of homozygosity calculated using PLINK. Samples are flagged as outliers if the observed and ancestry adjusted heterozygosity as well as the residuals from the models fit against runs of homozygosity are four standard deviations away from the mean. 166 samples from iPSYCH2012 and 98 samples from iPSYCH2015i failed this quality check and were excluded.

S1.3.2 Sample Duplication

A total of 121 samples were found to be genotyped more than once across the 26 waves in iPSYCH2012. Further, mapping sample identifiers to unique identifiers from the registers yielded 159 sample identifiers in iPSYCH2012 and 25 sample identifiers in iPSYCH2015i mapping to a non-unique identifier in the registry. Two samples from iPSYCH2012 were found to be genotyped again in iPSYCH2015i due to the randomness of ascertainment. In each case, the sample with lower missingness was retained. 6 samples in iPSYCH2012 and 1 sample in iPSYCH2015i were genotyped as part of the trios and were excluded.

Kinship analysis performed using KING[8] revealed three monozygotic twins in iPSYCH2012 and ten monozygotic twins in iPSYCH2015i. In each case, the case was retained and if both samples were cases, the sample with higher missingness was excluded.

S1.3.3 Sample Missingness:

Two samples from the iPSYCH2012 cohort were excluded for excessive missingness (> 5%).

This left us with 80,876 samples in iPSYCH2012, genotyped at 251,551 loci and 48,974 individuals in iPSYCH2015i, genotyped at 450,445 loci to be used as a backbone for haplotype estimation and missing data imputation.

S2. ANCESTRY COMPOSITION OF iPSYCH

Parental Birth Place	iPSYCH2012	iPSYCH2015i
Denmark	67044	41673
Denmark_Europe	2416	1591
Denmark_Scandinavia	1476	913
Europe	1169	785
Denmark_Unknown	829	543
MiddleEast	775	563
Asia	594	384
Asia_Denmark	581	292
Africa	473	277
Denmark_Greenland	435	284
Africa_Denmark	431	292
Denmark_NorthAmerica	363	234
Denmark_MiddleEast	354	216
Denmark_SouthAmerica	235	184
Scandinavia	109	67

Supplementary Table 2. Ancestry composition of iPSYCH by parental birthplace as obtained from the Danish Civil Registers[9]. Underscore delimited combinations indicate parents born in different regions.

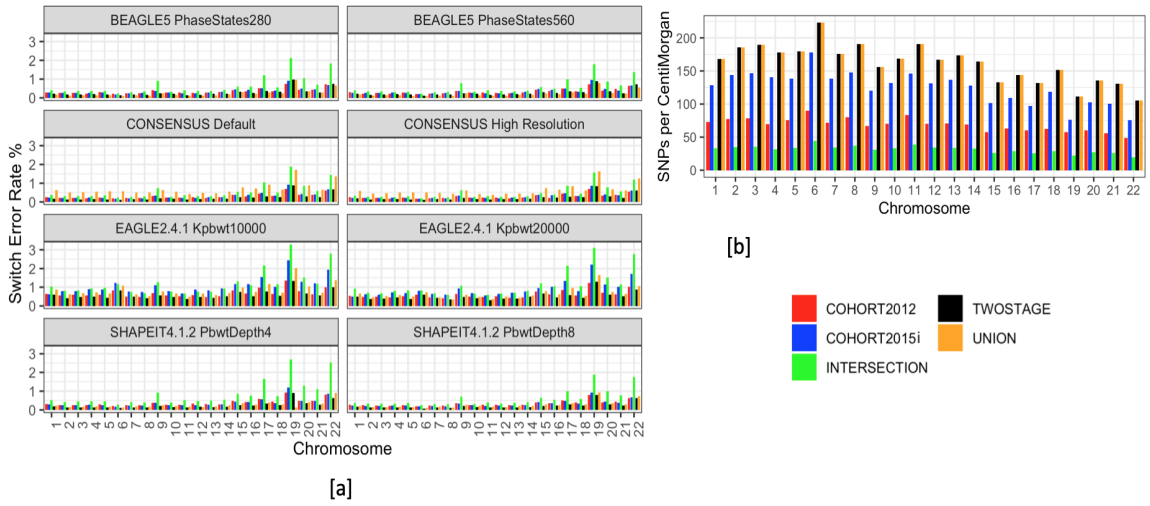
S3. SWITCH ERROR RATES:

STRATEGY	METHOD	PARAMETERS	SER%
COHORT2012	BEAGLE5	PhaseStates560	0.2991
COHORT2015i	BEAGLE5	PhaseStates560	0.3077
COHORT2012	BEAGLE5	PhaseStates280	0.308
COHORT2015i	BEAGLE5	PhaseStates280	0.3147
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	0.294
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	0.2745
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	0.3317
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	0.3133
COHORT2012	EAGLE2.4.1	Kpbwt20000	0.5504
COHORT2015i	EAGLE2.4.1	Kpbwt20000	0.7651
COHORT2012	EAGLE2.4.1	Kpbwt10000	0.6331
COHORT2015i	EAGLE2.4.1	Kpbwt10000	0.9454
COHORT2012	CONSENSUS	High Resolution	0.2534
COHORT2015i	CONSENSUS	High Resolution	0.2593
COHORT2012	CONSENSUS	Default	0.2624
COHORT2015i	CONSENSUS	Default	0.2689
INTERSECTION	BEAGLE5	PhaseStates560	0.4554
INTERSECTION	BEAGLE5	PhaseStates280	0.4995
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	0.4715
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	0.632
INTERSECTION	EAGLE2.4.1	Kpbwt20000	0.9204
INTERSECTION	EAGLE2.4.1	Kpbwt10000	1.0372

INTERSECTION	CONSENSUS	High Resolution	0.3773
INTERSECTION	CONSENSUS	Default	0.435
UNION	BEAGLE5	PhaseStates560	0.1743
UNION	BEAGLE5	PhaseStates280	0.1918
UNION	SHAPEIT4.1.2	PbwtDepth8	0.2192
UNION	SHAPEIT4.1.2	PbwtDepth4	0.2444
UNION	EAGLE2.4.1	Kpbwt20000	0.5831
UNION	EAGLE2.4.1	Kpbwt10000	0.7692
UNION	CONSENSUS	High Resolution	0.5513
UNION	CONSENSUS	Default	0.6137

Supplementary Table 3. Phasing accuracy as indicated by switch error rates obtained by comparing the mendelian transmission of phase to computationally estimated phase within 124 trio offspring for whom parental genotypes are known at heterozygous loci genotyped on both iPSYCH arrays.

The marker coverage from the iPSYCH genotyping arrays is not uniform across all chromosomes. As shown in supplementary figure 4, this leads to a variability in the accuracy of haplotype estimation by chromosome number.

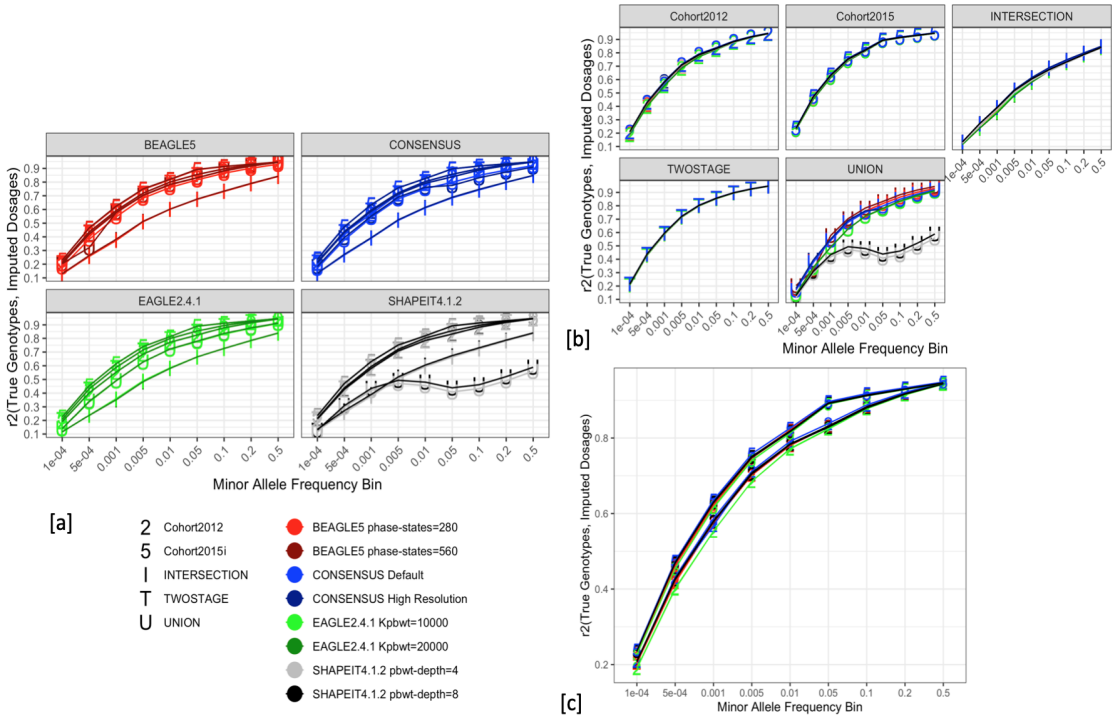


Supplementary figure 1. [a] SNP density across chromosomes within each data integration protocol. [b] Haplotype estimation accuracy as shown by switch error rates obtained from comparing computationally assigned phase to mendelian transmission in 124 trio offspring whose parental genotypes are known.

S4. IMPUTATION ACCURACY WITHIN PERSONAL GENOMES PROJECT - UK SAMPLES

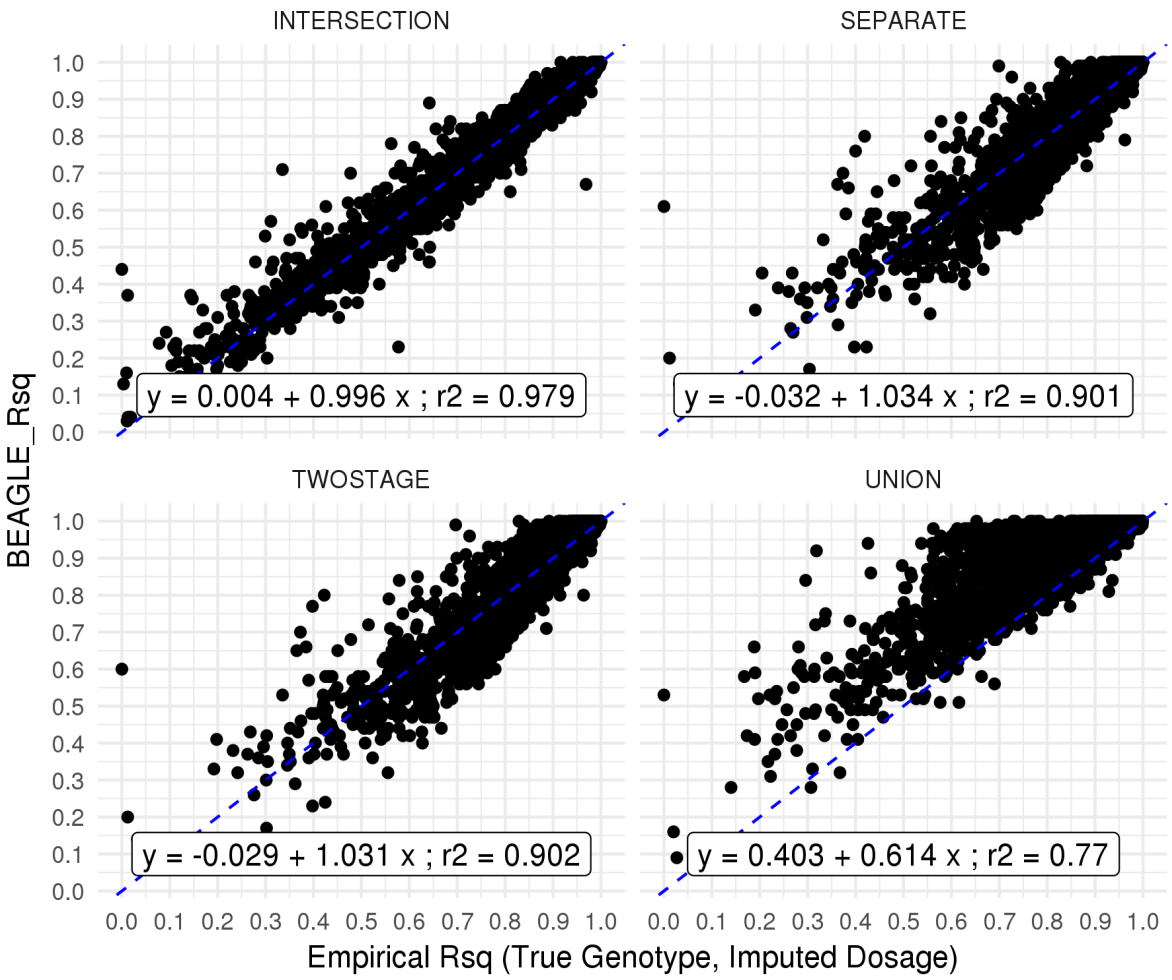
BAM files corresponding to 10 samples from the personal genomes project - UK[10] were downloaded from the EGA (study accession: PRJEB17529), sample accessions (SAMEA4545245, SAMEA4545246, SAMEA4545247, SAMEA4545248, SAMEA4545249, SAMEA4545250, SAMEA4545251, SAMEA4545252, SAMEA4545253, SAMEA4545254). Variant calling was performed using samtools mpileup and the samples were further down-sampled to each of the two iPSYCH genotyping arrays and added to cohorts arising from each data integration protocol prior to phasing and imputation. The accuracy of the imputation was calculated as the squared pearson correlation coefficient between the imputed dosages and variant calls at 6.5 million loci not genotyped on either iPSYCH array. The results as shown in supplementary figures 5a,b across minor allele frequency bins as ascertained from the HRCv1.1 haplotype reference panel show similar results to the results obtained by gauging the accuracy at the 10,000 SNPs masked prior to phasing. The accuracy of imputation appears to rely more on choice of data integration protocol than haplotype estimation tool. The haplotypes obtained from SHAPEIT4.1.2 in presence of high missingness introduced by the *union* protocol lead to inaccurate imputations.

A comparison of imputation accuracy between the two iPSYCH genotyping arrays as shown in supplementary figure 5c reveals that all tools yield more accurate imputations in the cohort generated using the denser Illumina global screening array v2.0 despite a relatively lesser sample size for haplotype estimation as compared to the cohort generated using the Infinium PsychChipv1.0 with less dense SNP information but a higher sample size.



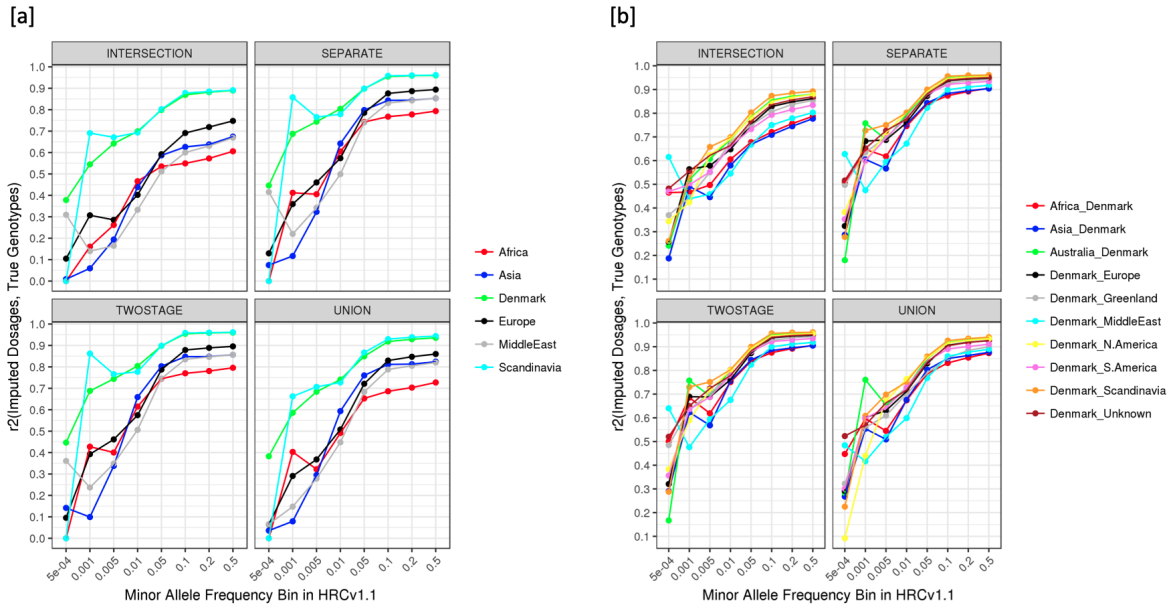
Supplementary Figure 2. Accuracy of imputation within the personal genomes project - UK whole genome sequenced samples, calculated as the squared pearson correlation coefficient between imputed dosages and true genotypes at loci not present on either iPSYCH genotyping array. [a] Grouped by choice of haplotype estimation tool. [b] Grouped by choice of data integration protocol. [c] Comparison between imputation accuracy obtained by using each iPSYCH genotyping array.

S5. RELIABILITY OF IMPUTATION QUALITY METRICS



Supplementary Figure 3. The relationship between empirical imputation accuracy, as measured by the squared pearson correlation coefficient true genotypes and imputed dosages at 10,000 masked SNPs, and BEAGLE r^2 within each data integration protocol. The plot shows the BEAGLE r2 is best calibrated for the imputations from the *intersection* protocol whereas it overestimates the accuracy of the imputation in the *union* protocol.

S6. IMPUTATION ACCURACY IN NON-EUROPEAN AND ADMIXED SAMPLES ACROSS DATA INTEGRATION PROTOCOLS



Supplementary Figure 4. Accuracy of imputation varies by parental origin. The attenuation in imputation accuracy within samples of non-European origin is further magnified by choice of data integration protocol. [a] Shows the accuracy of imputation within the 10,000 masked SNPs at different minor allele frequency bins within samples grouped by the birthplace of their parents according to the Danish civil registers across all four data integration protocols. [b] Shows the accuracy of imputation within the 10,000 masked SNPs within samples where at least one parent was born in Denmark.

S7. PGS ANALYSIS

PROTOCOL	Variance Explained by PGS from all alleles	Variance explained by PGS when the effect allele is the major allele in HRC	Variance explained by PGS when the effect allele is the minor allele in HRC	Instrument
TRUTH	0.4493	0.132	0.3176	True Genotype
TRUTH	0.4493	0.132	0.3176	True Genotype
SEPARATE	0.4194	0.1237	0.2961	Best Guess Genotype
SEPARATE	0.4258	0.126	0.3	Dosage
INTERSECTION	0.3765	0.1128	0.2641	Best Guess Genotype
INTERSECTION	0.3899	0.1158	0.275	Dosage
UNION	0.4092	0.1215	0.2894	Best Guess Genotype
UNION	0.4145	0.1229	0.2928	Dosage
TWOSTAGE	0.4199	0.1238	0.2967	Best Guess Genotype
TWOSTAGE	0.4259	0.126	0.3001	Dosage

Supplementary Table 4. Variance explained in a simulated continuous phenotype with a SNP heritability of 0.5 and the 10,000 masked SNPs as causal loci by true genotypes and imputed dosages, best guess genotypes across the four data integration protocols. The PGS is further divided into two components depending on if the effect allele is the major or minor allele in HRCv1.1.

PROTOCOL	Effect Allele	% Attenuation in variance explained by PGS as compared to true genotypes	Instrument
SEPARATE	Major Allele	6.28	Best Guess Genotype
SEPARATE	Minor Allele	6.76	Best Guess Genotype
SEPARATE	Major Allele	4.54	Dosage
SEPARATE	Minor Allele	5.54	Dosage
INTERSECTION	Major Allele	14.54	Best Guess Genotype
INTERSECTION	Minor Allele	16.84	Best Guess Genotype
INTERSECTION	Major Allele	12.27	Dosage
INTERSECTION	Minor Allele	13.41	Dosage
UNION	Major Allele	7.95	Best Guess Genotype
UNION	Minor Allele	8.87	Best Guess Genotype
UNION	Major Allele	6.89	Dosage
UNION	Minor Allele	7.81	Dosage
TWOSTAGE	Major Allele	6.21	Best Guess Genotype
TWOSTAGE	Minor Allele	6.58	Best Guess Genotype
TWOSTAGE	Major Allele	4.54	Dosage
TWOSTAGE	Minor Allele	5.51	Dosage
SEPARATE	ALL	6.65	Best Guess Genotype
SEPARATE	ALL	5.23	Dosage
INTERSECTION	ALL	16.2	Best Guess Genotype
INTERSECTION	ALL	13.22	Dosage
UNION	ALL	8.92	Best Guess Genotype

UNION	ALL	7.74	Dosage
TWOSTAGE	ALL	6.54	Best Guess Genotype
TWOSTAGE	ALL	5.21	Dosage

Supplementary Table 5. Attenuation in variance explained by PGS calculated using imputed dosages and best guess genotypes as compared to true genotypes across the four data integration protocols. The attenuation is further decomposed based on if the effect allele is the major or minor allele in HRCv1.1.

S8. BATCH EFFECTS

PROTOCOL	iPSYCH2012	iPSYCH2015i	R2	MAF	N	LAMBDA_GC
TWOSTAGE	Imputed	Imputed	0	0.01	9566	1.23
TWOSTAGE	Imputed	Imputed	0.3	0.01	9564	1.23
TWOSTAGE	Imputed	Imputed	0.6	0.01	9430	1.22
TWOSTAGE	Imputed	Imputed	0.8	0.01	8871	1.18
TWOSTAGE	Imputed	Imputed	0.9	0.01	7964	1.13
TWOSTAGE	Imputed	Imputed	0.95	0.01	6410	1.1
TWOSTAGE	Imputed	Imputed	1	0.01	1845	1.01
TWOSTAGE	Genotyped	Imputed	0	0.01	9543	1.18
TWOSTAGE	Genotyped	Imputed	0.3	0.01	9541	1.18
TWOSTAGE	Genotyped	Imputed	0.6	0.01	9407	1.17
TWOSTAGE	Genotyped	Imputed	0.8	0.01	8849	1.15
TWOSTAGE	Genotyped	Imputed	0.9	0.01	7945	1.11
TWOSTAGE	Genotyped	Imputed	0.95	0.01	6397	1.09
TWOSTAGE	Genotyped	Imputed	1	0.01	1841	1.01
TWOSTAGE	Imputed	Genotyped	0	0.01	9543	1.28

TWOSTAGE	Imputed	Genotyped	0.3	0.01	9541	1.28
TWOSTAGE	Imputed	Genotyped	0.6	0.01	9407	1.25
TWOSTAGE	Imputed	Genotyped	0.8	0.01	8849	1.21
TWOSTAGE	Imputed	Genotyped	0.9	0.01	7945	1.16
TWOSTAGE	Imputed	Genotyped	0.95	0.01	6397	1.09
TWOSTAGE	Imputed	Genotyped	1	0.01	1841	1
SEPARATE	Imputed	Imputed	0	0.01	9569	1.25
SEPARATE	Imputed	Imputed	0.3	0.01	9566	1.25
SEPARATE	Imputed	Imputed	0.6	0.01	9431	1.23
SEPARATE	Imputed	Imputed	0.8	0.01	8869	1.19
SEPARATE	Imputed	Imputed	0.9	0.01	7957	1.13
SEPARATE	Imputed	Imputed	0.95	0.01	6404	1.1
SEPARATE	Imputed	Imputed	1	0.01	1822	1
SEPARATE	Genotyped	Imputed	0	0.01	9546	1.18
SEPARATE	Genotyped	Imputed	0.3	0.01	9544	1.18
SEPARATE	Genotyped	Imputed	0.6	0.01	9497	1.17
SEPARATE	Genotyped	Imputed	0.8	0.01	9220	1.14
SEPARATE	Genotyped	Imputed	0.9	0.01	8565	1.1
SEPARATE	Genotyped	Imputed	0.95	0.01	7105	1.07
SEPARATE	Genotyped	Imputed	1	0.01	1747	1
SEPARATE	Imputed	Genotyped	0	0.01	9561	1.28
SEPARATE	Imputed	Genotyped	0.3	0.01	9559	1.28
SEPARATE	Imputed	Genotyped	0.6	0.01	9423	1.26
SEPARATE	Imputed	Genotyped	0.8	0.01	8858	1.21
SEPARATE	Imputed	Genotyped	0.9	0.01	7939	1.16

SEPARATE	Imputed	Genotyped	0.95	0.01	6384	1.09
SEPARATE	Imputed	Genotyped	1	0.01	1792	0.98
UNION	Imputed	Imputed	0	0.01	9571	2.81
UNION	Imputed	Imputed	0.3	0.01	9571	2.81
UNION	Imputed	Imputed	0.6	0.01	9536	2.8
UNION	Imputed	Imputed	0.8	0.01	9326	2.74
UNION	Imputed	Imputed	0.9	0.01	8799	2.62
UNION	Imputed	Imputed	0.95	0.01	7628	2.43
UNION	Imputed	Imputed	1	0.01	1478	1.93
UNION	Genotyped	Imputed	0	0.01	9548	1.77
UNION	Genotyped	Imputed	0.3	0.01	9548	1.77
UNION	Genotyped	Imputed	0.6	0.01	9513	1.76
UNION	Genotyped	Imputed	0.8	0.01	9304	1.74
UNION	Genotyped	Imputed	0.9	0.01	8779	1.7
UNION	Genotyped	Imputed	0.95	0.01	7612	1.68
UNION	Genotyped	Imputed	1	0.01	1476	1.59
UNION	Imputed	Genotyped	0	0.01	9548	2.13
UNION	Imputed	Genotyped	0.3	0.01	9548	2.13
UNION	Imputed	Genotyped	0.6	0.01	9513	2.11
UNION	Imputed	Genotyped	0.8	0.01	9304	2.04
UNION	Imputed	Genotyped	0.9	0.01	8779	1.94
UNION	Imputed	Genotyped	0.95	0.01	7612	1.75
UNION	Imputed	Genotyped	1	0.01	1476	1.23
INTERSECTION	Imputed	Imputed	0	0.01	9570	1
INTERSECTION	Imputed	Imputed	0.3	0.01	9499	1

INTERSECTION	Imputed	Imputed	0.6	0.01	8802	1
INTERSECTION	Imputed	Imputed	0.8	0.01	7183	1.01
INTERSECTION	Imputed	Imputed	0.9	0.01	5226	1
INTERSECTION	Imputed	Imputed	0.95	0.01	3096	0.97
INTERSECTION	Imputed	Imputed	1	0.01	421	0.89

Supplementary Table 6. Batch effects, as demonstrated by the inflation in test statistics when performing GWAS with genotyped and imputed dosages from the two iPSYCH cohorts at the 10,000 masked SNPs in controls of European origin with the genotyping array as the outcome.

* Supplementary tables 7 - 10 in appendix.

REFERENCES

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* vol. 83 132–5; author reply 135–9 (2008).
3. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
4. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
5. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
6. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568–584.e23 (2020).
7. Danecek, P., McCarthy, S., Li, H. & Others. bcftools—utilities for variant calling and manipulating vcfs and bcfs. (2015).
8. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
9. Pedersen, C. B. The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).
10. Chervova, O. *et al.* The Personal Genome Project-UK, an open access resource of human multi-omics data. *Sci Data* **6**, 257 (2019).
11. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993 (2011).

Supplementary Table 7 Imputation Accuracy in 10000 Masked SNPs

PROTOCOL	TOOL	MaxMAF	Rsq	N
SEPARATE	BEAGLE5	0.0005	0.4285	1038137
SEPARATE	BEAGLE5	0.001	0.6757	3504422
SEPARATE	BEAGLE5	0.005	0.7238	23876323
SEPARATE	BEAGLE5	0.01	0.7884	22186507
SEPARATE	BEAGLE5	0.05	0.8844	105975418
SEPARATE	BEAGLE5	0.1	0.9432	93394854
SEPARATE	BEAGLE5	0.2	0.9504	246249466
SEPARATE	BEAGLE5	0.5	0.9532	797422583
SEPARATE	SHAPEIT4.1.2	0.0005	0.4289	1038137
SEPARATE	SHAPEIT4.1.2	0.001	0.6767	3504422
SEPARATE	SHAPEIT4.1.2	0.005	0.7242	23876323
SEPARATE	SHAPEIT4.1.2	0.01	0.7895	22186507
SEPARATE	SHAPEIT4.1.2	0.05	0.8849	105975418
SEPARATE	SHAPEIT4.1.2	0.1	0.9437	93394854
SEPARATE	SHAPEIT4.1.2	0.2	0.9507	246249466
SEPARATE	SHAPEIT4.1.2	0.5	0.9534	797422583
SEPARATE	EAGLE2.4.1	0.0005	0.4276	1038137
SEPARATE	EAGLE2.4.1	0.001	0.6732	3504422
SEPARATE	EAGLE2.4.1	0.005	0.7203	23876323
SEPARATE	EAGLE2.4.1	0.01	0.785	22186507
SEPARATE	EAGLE2.4.1	0.05	0.8817	105975418
SEPARATE	EAGLE2.4.1	0.1	0.9417	93394854
SEPARATE	EAGLE2.4.1	0.2	0.9491	246249466
SEPARATE	EAGLE2.4.1	0.5	0.9521	797422583
SEPARATE	CONSENSUS	0.0005	0.4287	1038137
SEPARATE	CONSENSUS	0.001	0.6777	3504422
SEPARATE	CONSENSUS	0.005	0.7257	23876323
SEPARATE	CONSENSUS	0.01	0.7908	22186507
SEPARATE	CONSENSUS	0.05	0.8859	105975418
SEPARATE	CONSENSUS	0.1	0.9442	93394854
SEPARATE	CONSENSUS	0.2	0.9512	246249466
SEPARATE	CONSENSUS	0.5	0.9539	797422583
INTERSECTION	BEAGLE5	0.0005	0.3635	1038137
INTERSECTION	BEAGLE5	0.001	0.5343	3504422
INTERSECTION	BEAGLE5	0.005	0.6181	23876323
INTERSECTION	BEAGLE5	0.01	0.679	22186507
INTERSECTION	BEAGLE5	0.05	0.7733	105975418
INTERSECTION	BEAGLE5	0.1	0.8455	93394854
INTERSECTION	BEAGLE5	0.2	0.8647	246249466
INTERSECTION	BEAGLE5	0.5	0.8763	797422583
INTERSECTION	SHAPEIT4.1.2	0.0005	0.3615	1038137
INTERSECTION	SHAPEIT4.1.2	0.001	0.5336	3504422
INTERSECTION	SHAPEIT4.1.2	0.005	0.617	23876323
INTERSECTION	SHAPEIT4.1.2	0.01	0.6787	22186507
INTERSECTION	SHAPEIT4.1.2	0.05	0.773	105975418
INTERSECTION	SHAPEIT4.1.2	0.1	0.8452	93394854

INTERSECTION	SHAPEIT4.1.2	0.2	0.8644	246249466
INTERSECTION	SHAPEIT4.1.2	0.5	0.8758	797422583
INTERSECTION	EAGLE2.4.1	0.0005	0.3595	1038137
INTERSECTION	EAGLE2.4.1	0.001	0.5307	3504422
INTERSECTION	EAGLE2.4.1	0.005	0.6121	23876323
INTERSECTION	EAGLE2.4.1	0.01	0.672	22186507
INTERSECTION	EAGLE2.4.1	0.05	0.7669	105975418
INTERSECTION	EAGLE2.4.1	0.1	0.8409	93394854
INTERSECTION	EAGLE2.4.1	0.2	0.8609	246249466
INTERSECTION	EAGLE2.4.1	0.5	0.8731	797422583
INTERSECTION	CONSENSUS	0.0005	0.3636	1038137
INTERSECTION	CONSENSUS	0.001	0.5365	3504422
INTERSECTION	CONSENSUS	0.005	0.6202	23876323
INTERSECTION	CONSENSUS	0.01	0.6817	22186507
INTERSECTION	CONSENSUS	0.05	0.7756	105975418
INTERSECTION	CONSENSUS	0.1	0.8472	93394854
INTERSECTION	CONSENSUS	0.2	0.8661	246249466
INTERSECTION	CONSENSUS	0.5	0.8773	797422583
UNION	BEAGLE5	0.0005	0.3662	1038137
UNION	BEAGLE5	0.001	0.5757	3504422
UNION	BEAGLE5	0.005	0.6626	23876323
UNION	BEAGLE5	0.01	0.7254	22186507
UNION	BEAGLE5	0.05	0.8348	105975418
UNION	BEAGLE5	0.1	0.9058	93394854
UNION	BEAGLE5	0.2	0.9198	246249466
UNION	BEAGLE5	0.5	0.9282	797422583
UNION	SHAPEIT4.1.2	0.0005	0.0446	1038137
UNION	SHAPEIT4.1.2	0.001	0.0327	3504422
UNION	SHAPEIT4.1.2	0.005	0.0761	23876323
UNION	SHAPEIT4.1.2	0.01	0.1074	22186507
UNION	SHAPEIT4.1.2	0.05	0.1262	105975418
UNION	SHAPEIT4.1.2	0.1	0.2272	93394854
UNION	SHAPEIT4.1.2	0.2	0.3097	246249466
UNION	SHAPEIT4.1.2	0.5	0.3872	797422583
UNION	EAGLE2.4.1	0.0005	0.3621	1038137
UNION	EAGLE2.4.1	0.001	0.5843	3504422
UNION	EAGLE2.4.1	0.005	0.6594	23876323
UNION	EAGLE2.4.1	0.01	0.7219	22186507
UNION	EAGLE2.4.1	0.05	0.8294	105975418
UNION	EAGLE2.4.1	0.1	0.9044	93394854
UNION	EAGLE2.4.1	0.2	0.9165	246249466
UNION	EAGLE2.4.1	0.5	0.9221	797422583
UNION	CONSENSUS	0.0005	0.3359	1038137
UNION	CONSENSUS	0.001	0.5407	3504422
UNION	CONSENSUS	0.005	0.6289	23876323
UNION	CONSENSUS	0.01	0.6982	22186507
UNION	CONSENSUS	0.05	0.807	105975418
UNION	CONSENSUS	0.1	0.8898	93394854

UNION	CONSENSUS	0.2	0.9053	246249466
UNION	CONSENSUS	0.5	0.913	797422583
TWOSTAGE	BEAGLE5	0.0005	0.4287	1038137
TWOSTAGE	BEAGLE5	0.001	0.6771	3504422
TWOSTAGE	BEAGLE5	0.005	0.7237	23876323
TWOSTAGE	BEAGLE5	0.01	0.789	22186507
TWOSTAGE	BEAGLE5	0.05	0.8846	105975418
TWOSTAGE	BEAGLE5	0.1	0.9434	93394854
TWOSTAGE	BEAGLE5	0.2	0.9506	246249466
TWOSTAGE	BEAGLE5	0.5	0.9534	797422583
TWOSTAGE	SHAPEIT4.1.2	0.0005	0.4062	1038137
TWOSTAGE	SHAPEIT4.1.2	0.001	0.6582	3504422
TWOSTAGE	SHAPEIT4.1.2	0.005	0.7084	23876323
TWOSTAGE	SHAPEIT4.1.2	0.01	0.7724	22186507
TWOSTAGE	SHAPEIT4.1.2	0.05	0.8741	105975418
TWOSTAGE	SHAPEIT4.1.2	0.1	0.9374	93394854
TWOSTAGE	SHAPEIT4.1.2	0.2	0.9439	246249466
TWOSTAGE	SHAPEIT4.1.2	0.5	0.9471	797422583
TWOSTAGE	EAGLE2.4.1	0.0005	0.4045	1038137
TWOSTAGE	EAGLE2.4.1	0.001	0.6554	3504422
TWOSTAGE	EAGLE2.4.1	0.005	0.7064	23876323
TWOSTAGE	EAGLE2.4.1	0.01	0.7703	22186507
TWOSTAGE	EAGLE2.4.1	0.05	0.8728	105975418
TWOSTAGE	EAGLE2.4.1	0.1	0.9366	93394854
TWOSTAGE	EAGLE2.4.1	0.2	0.9433	246249466
TWOSTAGE	EAGLE2.4.1	0.5	0.9466	797422583
TWOSTAGE	CONSENSUS	0.0005	0.4065	1038137
TWOSTAGE	CONSENSUS	0.001	0.6582	3504422
TWOSTAGE	CONSENSUS	0.005	0.709	23876323
TWOSTAGE	CONSENSUS	0.01	0.7727	22186507
TWOSTAGE	CONSENSUS	0.05	0.8745	105975418
TWOSTAGE	CONSENSUS	0.1	0.9376	93394854
TWOSTAGE	CONSENSUS	0.2	0.9441	246249466
TWOSTAGE	CONSENSUS	0.5	0.9473	797422583

Supplementary Table 8 Imputation Accuracy in 10000 masked SNPs by Parental Origin

PROTOCOL	PARENTAL_ORIGIN	MaxMAF	N	Rsq
INTERSECTION	Africa	0.0005	6032	6.09E-07
INTERSECTION	Africa	0.001	20349	0.16030719
INTERSECTION	Africa	0.005	138657	0.26195191
INTERSECTION	Africa	0.01	128849	0.46609072
INTERSECTION	Africa	0.05	615087	0.53504127
INTERSECTION	Africa	0.1	542044	0.54925846
INTERSECTION	Africa	0.2	1429543	0.57237139
INTERSECTION	Africa	0.5	4630095	0.60591154
INTERSECTION	Africa_Denmark	0.0005	5828	0.46505777
INTERSECTION	Africa_Denmark	0.001	19669	0.46737206
INTERSECTION	Africa_Denmark	0.005	134034	0.49694809
INTERSECTION	Africa_Denmark	0.01	124549	0.60537197
INTERSECTION	Africa_Denmark	0.05	594853	0.67754478
INTERSECTION	Africa_Denmark	0.1	524212	0.72038086
INTERSECTION	Africa_Denmark	0.2	1382383	0.75626472
INTERSECTION	Africa_Denmark	0.5	4476392	0.78685435
INTERSECTION	Asia	0.0005	7934	0.00901555
INTERSECTION	Asia	0.001	26772	0.05937071
INTERSECTION	Asia	0.005	182419	0.19387779
INTERSECTION	Asia	0.01	169530	0.43874461
INTERSECTION	Asia	0.05	809427	0.58662443
INTERSECTION	Asia	0.1	713426	0.62608355
INTERSECTION	Asia	0.2	1881306	0.63785661
INTERSECTION	Asia	0.5	6093012	0.67405296
INTERSECTION	Asia_Denmark	0.0005	8955	0.18744613
INTERSECTION	Asia_Denmark	0.001	30226	0.49067785
INTERSECTION	Asia_Denmark	0.005	205941	0.4456199
INTERSECTION	Asia_Denmark	0.01	191375	0.57959463
INTERSECTION	Asia_Denmark	0.05	913819	0.66869566
INTERSECTION	Asia_Denmark	0.1	805380	0.70901911
INTERSECTION	Asia_Denmark	0.2	2123723	0.74508902
INTERSECTION	Asia_Denmark	0.5	6877066	0.77701049
INTERSECTION	Australia_Denmark	0.0005	991	0.24110562
INTERSECTION	Australia_Denmark	0.001	3348	0.51922924
INTERSECTION	Australia_Denmark	0.005	22811	0.60754839
INTERSECTION	Australia_Denmark	0.01	21191	0.68985084
INTERSECTION	Australia_Denmark	0.05	101208	0.78405567
INTERSECTION	Australia_Denmark	0.1	89181	0.8552119
INTERSECTION	Australia_Denmark	0.2	235181	0.87132345
INTERSECTION	Australia_Denmark	0.5	761379	0.87982476
INTERSECTION	Denmark	0.0005	879293	0.37809925
INTERSECTION	Denmark	0.001	2968255	0.54450221
INTERSECTION	Denmark	0.005	20223273	0.64207068
INTERSECTION	Denmark	0.01	18792891	0.69896837
INTERSECTION	Denmark	0.05	89767862	0.79868562
INTERSECTION	Denmark	0.1	79106797	0.86898652

INTERSECTION	Denmark	0.2	208582257	0.88185621
INTERSECTION	Denmark	0.5	675433806	0.88943077
INTERSECTION	Denmark_Europe	0.0005	32335	0.2565308
INTERSECTION	Denmark_Europe	0.001	109171	0.5635531
INTERSECTION	Denmark_Europe	0.005	743773	0.5783658
INTERSECTION	Denmark_Europe	0.01	691188	0.64787598
INTERSECTION	Denmark_Europe	0.05	3301430	0.75081937
INTERSECTION	Denmark_Europe	0.1	2909290	0.82832195
INTERSECTION	Denmark_Europe	0.2	7670946	0.84799367
INTERSECTION	Denmark_Europe	0.5	24839967	0.86143457
INTERSECTION	Denmark_Greenland	0.0005	5831	0.36975394
INTERSECTION	Denmark_Greenland	0.001	19679	0.44499104
INTERSECTION	Denmark_Greenland	0.005	134051	0.55061102
INTERSECTION	Denmark_Greenland	0.01	124570	0.67802368
INTERSECTION	Denmark_Greenland	0.05	595004	0.74864855
INTERSECTION	Denmark_Greenland	0.1	524340	0.80644251
INTERSECTION	Denmark_Greenland	0.2	1382577	0.8369938
INTERSECTION	Denmark_Greenland	0.5	4476896	0.85412418
INTERSECTION	Denmark_MiddleEast	0.0005	4614	0.61520707
INTERSECTION	Denmark_MiddleEast	0.001	15577	0.43858749
INTERSECTION	Denmark_MiddleEast	0.005	106113	0.45807116
INTERSECTION	Denmark_MiddleEast	0.01	98594	0.54546701
INTERSECTION	Denmark_MiddleEast	0.05	470948	0.66773151
INTERSECTION	Denmark_MiddleEast	0.1	415016	0.74993163
INTERSECTION	Denmark_MiddleEast	0.2	1094386	0.77869482
INTERSECTION	Denmark_MiddleEast	0.5	3544314	0.80264197
INTERSECTION	Denmark_Scandinavia	0.0005	19350	0.26068365
INTERSECTION	Denmark_Scandinavia	0.001	65308	0.52642897
INTERSECTION	Denmark_Scandinavia	0.005	445001	0.65779803
INTERSECTION	Denmark_Scandinavia	0.01	413516	0.69878443
INTERSECTION	Denmark_Scandinavia	0.05	1975193	0.80318282
INTERSECTION	Denmark_Scandinavia	0.1	1740548	0.87320763
INTERSECTION	Denmark_Scandinavia	0.2	4589569	0.88478664
INTERSECTION	Denmark_Scandinavia	0.5	14862134	0.8923238
INTERSECTION	Denmark_Unknown	0.0005	11031	0.4820796
INTERSECTION	Denmark_Unknown	0.001	37247	0.55519357
INTERSECTION	Denmark_Unknown	0.005	253760	0.62217457
INTERSECTION	Denmark_Unknown	0.01	235801	0.66394898
INTERSECTION	Denmark_Unknown	0.05	1126273	0.76575386
INTERSECTION	Denmark_Unknown	0.1	992613	0.8357535
INTERSECTION	Denmark_Unknown	0.2	2617213	0.85619788
INTERSECTION	Denmark_Unknown	0.5	8475381	0.8692286
INTERSECTION	Europe	0.0005	15782	0.10464072
INTERSECTION	Europe	0.001	53272	0.30696097
INTERSECTION	Europe	0.005	362957	0.2855984
INTERSECTION	Europe	0.01	337267	0.40234811
INTERSECTION	Europe	0.05	1610982	0.59216242
INTERSECTION	Europe	0.1	1419674	0.69115988

INTERSECTION	Europe	0.2	3743207	0.71899018
INTERSECTION	Europe	0.5	12122315	0.7476181
INTERSECTION	MiddleEast	0.0005	10742	0.30941238
INTERSECTION	MiddleEast	0.001	36245	0.13946181
INTERSECTION	MiddleEast	0.005	246965	0.16473836
INTERSECTION	MiddleEast	0.01	229499	0.33268566
INTERSECTION	MiddleEast	0.05	1096049	0.51142536
INTERSECTION	MiddleEast	0.1	965812	0.60014628
INTERSECTION	MiddleEast	0.2	2546894	0.63078355
INTERSECTION	MiddleEast	0.5	8247559	0.66916734
INTERSECTION	Scandinavia	0.0005	1423	8.71E-06
INTERSECTION	Scandinavia	0.001	4803	0.69078607
INTERSECTION	Scandinavia	0.005	32737	0.67117376
INTERSECTION	Scandinavia	0.01	30424	0.69360542
INTERSECTION	Scandinavia	0.05	145311	0.80201379
INTERSECTION	Scandinavia	0.1	128048	0.877935
INTERSECTION	Scandinavia	0.2	337693	0.88454127
INTERSECTION	Scandinavia	0.5	1093720	0.89102085
TWOSTAGE	Africa	0.0005	6032	4.19E-07
TWOSTAGE	Africa	0.001	20349	0.42748615
TWOSTAGE	Africa	0.005	138657	0.40046287
TWOSTAGE	Africa	0.01	128849	0.61468197
TWOSTAGE	Africa	0.05	615087	0.74444741
TWOSTAGE	Africa	0.1	542044	0.76999682
TWOSTAGE	Africa	0.2	1429543	0.78061733
TWOSTAGE	Africa	0.5	4630095	0.7957012
TWOSTAGE	Africa_Denmark	0.0005	5828	0.50057246
TWOSTAGE	Africa_Denmark	0.001	19669	0.68545918
TWOSTAGE	Africa_Denmark	0.005	134034	0.6188409
TWOSTAGE	Africa_Denmark	0.01	124549	0.75063635
TWOSTAGE	Africa_Denmark	0.05	594853	0.8389905
TWOSTAGE	Africa_Denmark	0.1	524212	0.87511303
TWOSTAGE	Africa_Denmark	0.2	1382383	0.89199754
TWOSTAGE	Africa_Denmark	0.5	4476392	0.90553351
TWOSTAGE	Asia	0.0005	7934	0.14163771
TWOSTAGE	Asia	0.001	26772	0.0989655
TWOSTAGE	Asia	0.005	182419	0.33783412
TWOSTAGE	Asia	0.01	169530	0.65885478
TWOSTAGE	Asia	0.05	809427	0.80304537
TWOSTAGE	Asia	0.1	713426	0.84755784
TWOSTAGE	Asia	0.2	1881306	0.84790798
TWOSTAGE	Asia	0.5	6093012	0.85660654
TWOSTAGE	Asia_Denmark	0.0005	8955	0.2905613
TWOSTAGE	Asia_Denmark	0.001	30226	0.62322719
TWOSTAGE	Asia_Denmark	0.005	205941	0.56825756
TWOSTAGE	Asia_Denmark	0.01	191375	0.75527665
TWOSTAGE	Asia_Denmark	0.05	913819	0.84412766
TWOSTAGE	Asia_Denmark	0.1	805380	0.883757

TWOSTAGE	Asia_Denmark	0.2	2123723	0.89506498
TWOSTAGE	Asia_Denmark	0.5	6877066	0.90458276
TWOSTAGE	Australia_Denmark	0.0005	991	0.16688212
TWOSTAGE	Australia_Denmark	0.001	3348	0.75712812
TWOSTAGE	Australia_Denmark	0.005	22811	0.69210428
TWOSTAGE	Australia_Denmark	0.01	21191	0.80183852
TWOSTAGE	Australia_Denmark	0.05	101208	0.88893512
TWOSTAGE	Australia_Denmark	0.1	89181	0.95033305
TWOSTAGE	Australia_Denmark	0.2	235181	0.95348309
TWOSTAGE	Australia_Denmark	0.5	761379	0.95606008
TWOSTAGE	Denmark	0.0005	879293	0.44668727
TWOSTAGE	Denmark	0.001	2968255	0.6878208
TWOSTAGE	Denmark	0.005	20223273	0.74390259
TWOSTAGE	Denmark	0.01	18792891	0.80406288
TWOSTAGE	Denmark	0.05	89767862	0.89779906
TWOSTAGE	Denmark	0.1	79106797	0.95467082
TWOSTAGE	Denmark	0.2	208582257	0.95870673
TWOSTAGE	Denmark	0.5	675433806	0.95984355
TWOSTAGE	Denmark_Europe	0.0005	32335	0.32068606
TWOSTAGE	Denmark_Europe	0.001	109171	0.68904085
TWOSTAGE	Denmark_Europe	0.005	743773	0.68793612
TWOSTAGE	Denmark_Europe	0.01	691188	0.76119305
TWOSTAGE	Denmark_Europe	0.05	3301430	0.87308532
TWOSTAGE	Denmark_Europe	0.1	2909290	0.93725962
TWOSTAGE	Denmark_Europe	0.2	7670946	0.9447744
TWOSTAGE	Denmark_Europe	0.5	24839967	0.9475665
TWOSTAGE	Denmark_Greenland	0.0005	5831	0.48489579
TWOSTAGE	Denmark_Greenland	0.001	19679	0.63455949
TWOSTAGE	Denmark_Greenland	0.005	134051	0.70369085
TWOSTAGE	Denmark_Greenland	0.01	124570	0.77819967
TWOSTAGE	Denmark_Greenland	0.05	595004	0.88068621
TWOSTAGE	Denmark_Greenland	0.1	524340	0.92692876
TWOSTAGE	Denmark_Greenland	0.2	1382577	0.93782391
TWOSTAGE	Denmark_Greenland	0.5	4476896	0.94212236
TWOSTAGE	Denmark_MiddleEast	0.0005	4614	0.64018375
TWOSTAGE	Denmark_MiddleEast	0.001	15577	0.47600988
TWOSTAGE	Denmark_MiddleEast	0.005	106113	0.59413849
TWOSTAGE	Denmark_MiddleEast	0.01	98594	0.67520708
TWOSTAGE	Denmark_MiddleEast	0.05	470948	0.82416852
TWOSTAGE	Denmark_MiddleEast	0.1	415016	0.8996372
TWOSTAGE	Denmark_MiddleEast	0.2	1094386	0.91085959
TWOSTAGE	Denmark_MiddleEast	0.5	3544314	0.91805345
TWOSTAGE	Denmark_Scandinavia	0.0005	19350	0.28763423
TWOSTAGE	Denmark_Scandinavia	0.001	65308	0.72974312
TWOSTAGE	Denmark_Scandinavia	0.005	445001	0.75121138
TWOSTAGE	Denmark_Scandinavia	0.01	413516	0.8027122
TWOSTAGE	Denmark_Scandinavia	0.05	1975193	0.89961479
TWOSTAGE	Denmark_Scandinavia	0.1	1740548	0.95611683

TWOSTAGE	Denmark_Scandinavia	0.2	4589569	0.95964281
TWOSTAGE	Denmark_Scandinavia	0.5	14862134	0.96090096
TWOSTAGE	Denmark_Unknown	0.0005	11031	0.52032379
TWOSTAGE	Denmark_Unknown	0.001	37247	0.64968751
TWOSTAGE	Denmark_Unknown	0.005	253760	0.72381958
TWOSTAGE	Denmark_Unknown	0.01	235801	0.7784379
TWOSTAGE	Denmark_Unknown	0.05	1126273	0.8815661
TWOSTAGE	Denmark_Unknown	0.1	992613	0.93973679
TWOSTAGE	Denmark_Unknown	0.2	2617213	0.94689566
TWOSTAGE	Denmark_Unknown	0.5	8475381	0.95038044
TWOSTAGE	Europe	0.0005	15782	0.09525952
TWOSTAGE	Europe	0.001	53272	0.39250569
TWOSTAGE	Europe	0.005	362957	0.46147827
TWOSTAGE	Europe	0.01	337267	0.57451341
TWOSTAGE	Europe	0.05	1610982	0.78731779
TWOSTAGE	Europe	0.1	1419674	0.87860996
TWOSTAGE	Europe	0.2	3743207	0.88876741
TWOSTAGE	Europe	0.5	12122315	0.89570813
TWOSTAGE	MiddleEast	0.0005	10742	0.36096327
TWOSTAGE	MiddleEast	0.001	36245	0.2369339
TWOSTAGE	MiddleEast	0.005	246965	0.34817635
TWOSTAGE	MiddleEast	0.01	229499	0.5060451
TWOSTAGE	MiddleEast	0.05	1096049	0.7454003
TWOSTAGE	MiddleEast	0.1	965812	0.83521654
TWOSTAGE	MiddleEast	0.2	2546894	0.84656898
TWOSTAGE	MiddleEast	0.5	8247559	0.85677072
TWOSTAGE	Scandinavia	0.0005	1423	3.53E-06
TWOSTAGE	Scandinavia	0.001	4803	0.86203132
TWOSTAGE	Scandinavia	0.005	32737	0.76573199
TWOSTAGE	Scandinavia	0.01	30424	0.77696957
TWOSTAGE	Scandinavia	0.05	145311	0.89844825
TWOSTAGE	Scandinavia	0.1	128048	0.95811383
TWOSTAGE	Scandinavia	0.2	337693	0.96006611
TWOSTAGE	Scandinavia	0.5	1093720	0.96157327
UNION	Africa	0.0005	6032	4.81E-07
UNION	Africa	0.001	20349	0.40308178
UNION	Africa	0.005	138657	0.32228416
UNION	Africa	0.01	128849	0.49127392
UNION	Africa	0.05	615087	0.65287517
UNION	Africa	0.1	542044	0.68612942
UNION	Africa	0.2	1429543	0.70330285
UNION	Africa	0.5	4630095	0.7271927
UNION	Africa_Denmark	0.0005	5828	0.44740167
UNION	Africa_Denmark	0.001	19669	0.59642546
UNION	Africa_Denmark	0.005	134034	0.54482502
UNION	Africa_Denmark	0.01	124549	0.67367306
UNION	Africa_Denmark	0.05	594853	0.78586657
UNION	Africa_Denmark	0.1	524212	0.83137902

UNION	Africa_Denmark	0.2	1382383	0.85385761
UNION	Africa_Denmark	0.5	4476392	0.87287693
UNION	Asia	0.0005	7934	0.03581937
UNION	Asia	0.001	26772	0.0789579
UNION	Asia	0.005	182419	0.29495255
UNION	Asia	0.01	169530	0.59348909
UNION	Asia	0.05	809427	0.76010143
UNION	Asia	0.1	713426	0.81155035
UNION	Asia	0.2	1881306	0.81278418
UNION	Asia	0.5	6093012	0.8241183
UNION	Asia_Denmark	0.0005	8955	0.26760448
UNION	Asia_Denmark	0.001	30226	0.55378236
UNION	Asia_Denmark	0.005	205941	0.50902757
UNION	Asia_Denmark	0.01	191375	0.67600041
UNION	Asia_Denmark	0.05	913819	0.80269352
UNION	Asia_Denmark	0.1	805380	0.85002726
UNION	Asia_Denmark	0.2	2123723	0.86291631
UNION	Asia_Denmark	0.5	6877066	0.87632557
UNION	Australia_Denmark	0.0005	991	0.28288502
UNION	Australia_Denmark	0.001	3348	0.76038126
UNION	Australia_Denmark	0.005	22811	0.6537551
UNION	Australia_Denmark	0.01	21191	0.72654333
UNION	Australia_Denmark	0.05	101208	0.8491628
UNION	Australia_Denmark	0.1	89181	0.91687055
UNION	Australia_Denmark	0.2	235181	0.92854245
UNION	Australia_Denmark	0.5	761379	0.93442532
UNION	Denmark	0.0005	879293	0.38268637
UNION	Denmark	0.001	2968255	0.58561835
UNION	Denmark	0.005	20223273	0.68407155
UNION	Denmark	0.01	18792891	0.74075751
UNION	Denmark	0.05	89767862	0.84923587
UNION	Denmark	0.1	79106797	0.91815653
UNION	Denmark	0.2	208582257	0.92901034
UNION	Denmark	0.5	675433806	0.9354558
UNION	Denmark_Europe	0.0005	32335	0.29133231
UNION	Denmark_Europe	0.001	109171	0.60092326
UNION	Denmark_Europe	0.005	743773	0.63044911
UNION	Denmark_Europe	0.01	691188	0.71166318
UNION	Denmark_Europe	0.05	3301430	0.8296465
UNION	Denmark_Europe	0.1	2909290	0.90612705
UNION	Denmark_Europe	0.2	7670946	0.91878486
UNION	Denmark_Europe	0.5	24839967	0.92558096
UNION	Denmark_Greenland	0.0005	5831	0.32200235
UNION	Denmark_Greenland	0.001	19679	0.55908142
UNION	Denmark_Greenland	0.005	134051	0.60924003
UNION	Denmark_Greenland	0.01	124570	0.70164843
UNION	Denmark_Greenland	0.05	595004	0.7886031
UNION	Denmark_Greenland	0.1	524340	0.85305822

UNION	Denmark_Greenland	0.2	1382577	0.88443677
UNION	Denmark_Greenland	0.5	4476896	0.89949822
UNION	Denmark_MiddleEast	0.0005	4614	0.48407154
UNION	Denmark_MiddleEast	0.001	15577	0.4166135
UNION	Denmark_MiddleEast	0.005	106113	0.51918031
UNION	Denmark_MiddleEast	0.01	98594	0.5985507
UNION	Denmark_MiddleEast	0.05	470948	0.76776129
UNION	Denmark_MiddleEast	0.1	415016	0.85921169
UNION	Denmark_MiddleEast	0.2	1094386	0.87698663
UNION	Denmark_MiddleEast	0.5	3544314	0.88919928
UNION	Denmark_Scandinavia	0.0005	19350	0.22494006
UNION	Denmark_Scandinavia	0.001	65308	0.6081232
UNION	Denmark_Scandinavia	0.005	445001	0.69795697
UNION	Denmark_Scandinavia	0.01	413516	0.7512094
UNION	Denmark_Scandinavia	0.05	1975193	0.85873252
UNION	Denmark_Scandinavia	0.1	1740548	0.92547786
UNION	Denmark_Scandinavia	0.2	4589569	0.93438441
UNION	Denmark_Scandinavia	0.5	14862134	0.94016794
UNION	Denmark_Unknown	0.0005	11031	0.52296569
UNION	Denmark_Unknown	0.001	37247	0.56928451
UNION	Denmark_Unknown	0.005	253760	0.66318775
UNION	Denmark_Unknown	0.01	235801	0.71691697
UNION	Denmark_Unknown	0.05	1126273	0.83449417
UNION	Denmark_Unknown	0.1	992613	0.90402985
UNION	Denmark_Unknown	0.2	2617213	0.9179257
UNION	Denmark_Unknown	0.5	8475381	0.92592929
UNION	Europe	0.0005	15782	0.06459057
UNION	Europe	0.001	53272	0.29082941
UNION	Europe	0.005	362957	0.36791837
UNION	Europe	0.01	337267	0.50776559
UNION	Europe	0.05	1610982	0.72107157
UNION	Europe	0.1	1419674	0.82961055
UNION	Europe	0.2	3743207	0.84715192
UNION	Europe	0.5	12122315	0.86011863
UNION	MiddleEast	0.0005	10742	0.06572217
UNION	MiddleEast	0.001	36245	0.14752533
UNION	MiddleEast	0.005	246965	0.27837425
UNION	MiddleEast	0.01	229499	0.44808081
UNION	MiddleEast	0.05	1096049	0.68342951
UNION	MiddleEast	0.1	965812	0.7878935
UNION	MiddleEast	0.2	2546894	0.80496181
UNION	MiddleEast	0.5	8247559	0.82026525
UNION	Scandinavia	0.0005	1423	3.83E-06
UNION	Scandinavia	0.001	4803	0.66243685
UNION	Scandinavia	0.005	32737	0.70661864
UNION	Scandinavia	0.01	30424	0.7274931
UNION	Scandinavia	0.05	145311	0.86684416
UNION	Scandinavia	0.1	128048	0.92936245

UNION	Scandinavia	0.2	337693	0.93848866
UNION	Scandinavia	0.5	1093720	0.94371261
SEPARATE	Africa_Denmark	0.0005	5828	0.50742959
SEPARATE	Africa_Denmark	0.001	19669	0.64705921
SEPARATE	Africa_Denmark	0.005	134034	0.61773233
SEPARATE	Africa_Denmark	0.01	124549	0.74576713
SEPARATE	Africa_Denmark	0.05	594853	0.838202
SEPARATE	Africa_Denmark	0.1	524212	0.87412551
SEPARATE	Africa_Denmark	0.2	1382383	0.89172058
SEPARATE	Africa_Denmark	0.5	4476392	0.9052777
SEPARATE	Africa	0.0005	6032	5.00E-07
SEPARATE	Africa	0.001	20349	0.41205244
SEPARATE	Africa	0.005	138657	0.40542049
SEPARATE	Africa	0.01	128849	0.6047978
SEPARATE	Africa	0.05	615087	0.74269348
SEPARATE	Africa	0.1	542044	0.7673323
SEPARATE	Africa	0.2	1429543	0.77781725
SEPARATE	Africa	0.5	4630095	0.79348904
SEPARATE	Asia_Denmark	0.0005	8955	0.28691599
SEPARATE	Asia_Denmark	0.001	30226	0.60519988
SEPARATE	Asia_Denmark	0.005	205941	0.56647391
SEPARATE	Asia_Denmark	0.01	191375	0.75288145
SEPARATE	Asia_Denmark	0.05	913819	0.84340504
SEPARATE	Asia_Denmark	0.1	805380	0.88310936
SEPARATE	Asia_Denmark	0.2	2123723	0.89457304
SEPARATE	Asia_Denmark	0.5	6877066	0.90420231
SEPARATE	Asia	0.0005	7934	0.07465744
SEPARATE	Asia	0.001	26772	0.11686693
SEPARATE	Asia	0.005	182419	0.32310994
SEPARATE	Asia	0.01	169530	0.64193667
SEPARATE	Asia	0.05	809427	0.79807726
SEPARATE	Asia	0.1	713426	0.84323992
SEPARATE	Asia	0.2	1881306	0.84382174
SEPARATE	Asia	0.5	6093012	0.85326578
SEPARATE	Australia_Denmark	0.0005	991	0.17990776
SEPARATE	Australia_Denmark	0.001	3348	0.75756928
SEPARATE	Australia_Denmark	0.005	22811	0.68999896
SEPARATE	Australia_Denmark	0.01	21191	0.79430386
SEPARATE	Australia_Denmark	0.05	101208	0.88826464
SEPARATE	Australia_Denmark	0.1	89181	0.94997251
SEPARATE	Australia_Denmark	0.2	235181	0.95328
SEPARATE	Australia_Denmark	0.5	761379	0.95598331
SEPARATE	Denmark_Europe	0.0005	32335	0.32409329
SEPARATE	Denmark_Europe	0.001	109171	0.68241515
SEPARATE	Denmark_Europe	0.005	743773	0.68723466
SEPARATE	Denmark_Europe	0.01	691188	0.76034052
SEPARATE	Denmark_Europe	0.05	3301430	0.87266295
SEPARATE	Denmark_Europe	0.1	2909290	0.93682829

SEPARATE	Denmark_Europe	0.2	7670946	0.94463044
SEPARATE	Denmark_Europe	0.5	24839967	0.9473651
SEPARATE	Denmark_Greenland	0.0005	5831	0.49751443
SEPARATE	Denmark_Greenland	0.001	19679	0.63469501
SEPARATE	Denmark_Greenland	0.005	134051	0.69822324
SEPARATE	Denmark_Greenland	0.01	124570	0.77800477
SEPARATE	Denmark_Greenland	0.05	595004	0.88094403
SEPARATE	Denmark_Greenland	0.1	524340	0.92694843
SEPARATE	Denmark_Greenland	0.2	1382577	0.93778256
SEPARATE	Denmark_Greenland	0.5	4476896	0.9421619
SEPARATE	Denmark	0.0005	879293	0.44603917
SEPARATE	Denmark	0.001	2968255	0.68716225
SEPARATE	Denmark	0.005	20223273	0.74414975
SEPARATE	Denmark	0.01	18792891	0.80389063
SEPARATE	Denmark	0.05	89767862	0.89787361
SEPARATE	Denmark	0.1	79106797	0.9547053
SEPARATE	Denmark	0.2	208582257	0.95871776
SEPARATE	Denmark	0.5	675433806	0.95985929
SEPARATE	Denmark_MiddleEast	0.0005	4614	0.62769018
SEPARATE	Denmark_MiddleEast	0.001	15577	0.47516875
SEPARATE	Denmark_MiddleEast	0.005	106113	0.59475633
SEPARATE	Denmark_MiddleEast	0.01	98594	0.67175124
SEPARATE	Denmark_MiddleEast	0.05	470948	0.82296206
SEPARATE	Denmark_MiddleEast	0.1	415016	0.89888081
SEPARATE	Denmark_MiddleEast	0.2	1094386	0.91042025
SEPARATE	Denmark_MiddleEast	0.5	3544314	0.91765185
SEPARATE	Denmark_N.America	0.0005	4812	0.3818871
SEPARATE	Denmark_N.America	0.001	16249	0.5993047
SEPARATE	Denmark_N.America	0.005	110686	0.71797744
SEPARATE	Denmark_N.America	0.01	102855	0.80237443
SEPARATE	Denmark_N.America	0.05	491290	0.89063975
SEPARATE	Denmark_N.America	0.1	432963	0.94708279
SEPARATE	Denmark_N.America	0.2	1141608	0.95218716
SEPARATE	Denmark_N.America	0.5	3696887	0.95486156
SEPARATE	Denmark_S.America	0.0005	3380	0.35284301
SEPARATE	Denmark_S.America	0.001	11416	0.59953464
SEPARATE	Denmark_S.America	0.005	77777	0.69511687
SEPARATE	Denmark_S.America	0.01	72281	0.78076307
SEPARATE	Denmark_S.America	0.05	345165	0.87895121
SEPARATE	Denmark_S.America	0.1	304198	0.92146827
SEPARATE	Denmark_S.America	0.2	802049	0.92838511
SEPARATE	Denmark_S.America	0.5	2597145	0.93427745
SEPARATE	Denmark_Scandinavia	0.0005	19350	0.27776129
SEPARATE	Denmark_Scandinavia	0.001	65308	0.7270291
SEPARATE	Denmark_Scandinavia	0.005	445001	0.75010614
SEPARATE	Denmark_Scandinavia	0.01	413516	0.80279475
SEPARATE	Denmark_Scandinavia	0.05	1975193	0.89980055
SEPARATE	Denmark_Scandinavia	0.1	1740548	0.9560483

SEPARATE	Denmark_Scandinavia	0.2	4589569	0.95959694
SEPARATE	Denmark_Scandinavia	0.5	14862134	0.96087843
SEPARATE	Denmark_Unknown	0.0005	11031	0.51614155
SEPARATE	Denmark_Unknown	0.001	37247	0.65145099
SEPARATE	Denmark_Unknown	0.005	253760	0.72615196
SEPARATE	Denmark_Unknown	0.01	235801	0.77577029
SEPARATE	Denmark_Unknown	0.05	1126273	0.88146762
SEPARATE	Denmark_Unknown	0.1	992613	0.93961439
SEPARATE	Denmark_Unknown	0.2	2617213	0.94683655
SEPARATE	Denmark_Unknown	0.5	8475381	0.9503559
SEPARATE	Europe	0.0005	15782	0.12938487
SEPARATE	Europe	0.001	53272	0.35934991
SEPARATE	Europe	0.005	362957	0.4603522
SEPARATE	Europe	0.01	337267	0.57341817
SEPARATE	Europe	0.05	1610982	0.78553892
SEPARATE	Europe	0.1	1419674	0.87593694
SEPARATE	Europe	0.2	3743207	0.88653529
SEPARATE	Europe	0.5	12122315	0.89393043
SEPARATE	MiddleEast	0.0005	10742	0.41555246
SEPARATE	MiddleEast	0.001	36245	0.21987386
SEPARATE	MiddleEast	0.005	246965	0.34225789
SEPARATE	MiddleEast	0.01	229499	0.49888925
SEPARATE	MiddleEast	0.05	1096049	0.73999209
SEPARATE	MiddleEast	0.1	965812	0.83086451
SEPARATE	MiddleEast	0.2	2546894	0.84283819
SEPARATE	MiddleEast	0.5	8247559	0.85342015
SEPARATE	Scandinavia	0.0005	1423	4.11E-06
SEPARATE	Scandinavia	0.001	4803	0.85763742
SEPARATE	Scandinavia	0.005	32737	0.76522181
SEPARATE	Scandinavia	0.01	30424	0.77893551
SEPARATE	Scandinavia	0.05	145311	0.89815199
SEPARATE	Scandinavia	0.1	128048	0.95793936
SEPARATE	Scandinavia	0.2	337693	0.95988173
SEPARATE	Scandinavia	0.5	1093720	0.96136981
INTERSECTION	Denmark_N.America	0.0005	4812	0.34414116
INTERSECTION	Denmark_N.America	0.001	16249	0.42425492
INTERSECTION	Denmark_N.America	0.005	110686	0.62313856
INTERSECTION	Denmark_N.America	0.01	102855	0.69248137
INTERSECTION	Denmark_N.America	0.05	491290	0.78154797
INTERSECTION	Denmark_N.America	0.1	432963	0.85035476
INTERSECTION	Denmark_N.America	0.2	1141608	0.86803176
INTERSECTION	Denmark_N.America	0.5	3696887	0.87892389
INTERSECTION	Denmark_S.America	0.0005	3380	0.46950989
INTERSECTION	Denmark_S.America	0.001	11416	0.49807642
INTERSECTION	Denmark_S.America	0.005	77777	0.55199021
INTERSECTION	Denmark_S.America	0.01	72281	0.66110762
INTERSECTION	Denmark_S.America	0.05	345165	0.73287675
INTERSECTION	Denmark_S.America	0.1	304198	0.79293023

INTERSECTION	Denmark_S.America	0.2	802049	0.81564301
INTERSECTION	Denmark_S.America	0.5	2597145	0.83417198
TWOSTAGE	Denmark_N.America	0.0005	4812	0.38296828
TWOSTAGE	Denmark_N.America	0.001	16249	0.59001015
TWOSTAGE	Denmark_N.America	0.005	110686	0.71893011
TWOSTAGE	Denmark_N.America	0.01	102855	0.79988518
TWOSTAGE	Denmark_N.America	0.05	491290	0.89068943
TWOSTAGE	Denmark_N.America	0.1	432963	0.94703359
TWOSTAGE	Denmark_N.America	0.2	1141608	0.95219686
TWOSTAGE	Denmark_N.America	0.5	3696887	0.95485467
TWOSTAGE	Denmark_S.America	0.0005	3380	0.35648049
TWOSTAGE	Denmark_S.America	0.001	11416	0.63444578
TWOSTAGE	Denmark_S.America	0.005	77777	0.68836054
TWOSTAGE	Denmark_S.America	0.01	72281	0.77918853
TWOSTAGE	Denmark_S.America	0.05	345165	0.87910225
TWOSTAGE	Denmark_S.America	0.1	304198	0.92168711
TWOSTAGE	Denmark_S.America	0.2	802049	0.92853222
TWOSTAGE	Denmark_S.America	0.5	2597145	0.93445854
UNION	Denmark_N.America	0.0005	4812	0.09177288
UNION	Denmark_N.America	0.001	16249	0.44000467
UNION	Denmark_N.America	0.005	110686	0.6714314
UNION	Denmark_N.America	0.01	102855	0.76306846
UNION	Denmark_N.America	0.05	491290	0.85137207
UNION	Denmark_N.America	0.1	432963	0.91564426
UNION	Denmark_N.America	0.2	1141608	0.92730337
UNION	Denmark_N.America	0.5	3696887	0.93363323
UNION	Denmark_S.America	0.0005	3380	0.30035362
UNION	Denmark_S.America	0.001	11416	0.59820306
UNION	Denmark_S.America	0.005	77777	0.64293814
UNION	Denmark_S.America	0.01	72281	0.72774582
UNION	Denmark_S.America	0.05	345165	0.83820384
UNION	Denmark_S.America	0.1	304198	0.89025411
UNION	Denmark_S.America	0.2	802049	0.90092966
UNION	Denmark_S.America	0.5	2597145	0.91031057

Supplementary Table 9 Imputation Accuracy in PGP UK Samples

PROTOCOL	TOOL	PRAMETER	MaxMAF	Rsq	N
Cohort2012	BEAGLE5	phase-states=560	0.0001	0.2038	86071
Cohort2012	BEAGLE5	phase-states=560	0.0005	0.4209	1181521
Cohort2012	BEAGLE5	phase-states=560	0.001	0.5786	1186016
Cohort2012	BEAGLE5	phase-states=560	0.005	0.7039	5460721
Cohort2012	BEAGLE5	phase-states=560	0.01	0.7834	2962562
Cohort2012	BEAGLE5	phase-states=560	0.05	0.832	9267817
Cohort2012	BEAGLE5	phase-states=560	0.1	0.8828	8469696
Cohort2012	BEAGLE5	phase-states=560	0.2	0.9181	129077490
Cohort2012	BEAGLE5	phase-states=560	0.5	0.9445	23504726
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.0001	0.2103	86071
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.0005	0.4284	1181521
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.001	0.5806	1186016
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.005	0.7082	5460721
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.01	0.7847	2962562
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.05	0.8319	9267817
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.1	0.8832	8469696
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.2	0.9179	129077490
Cohort2012	SHAPEIT4.1.2	pbwt-depth=8	0.5	0.9442	23504726
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.0001	0.191	86071
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.0005	0.4022	1181521
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.001	0.5532	1186016
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.005	0.6851	5460721
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.01	0.772	2962562
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.05	0.8249	9267817
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.1	0.8787	8469696
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.2	0.9153	129077490
Cohort2012	EAGLE2.4.1	Kpbwt=20000	0.5	0.9429	23504726
Cohort2012	CONSENSUS	High Resolution	0.0001	0.2085	86071
Cohort2012	CONSENSUS	High Resolution	0.0005	0.4324	1181521
Cohort2012	CONSENSUS	High Resolution	0.001	0.5873	1186016
Cohort2012	CONSENSUS	High Resolution	0.005	0.7126	5460721
Cohort2012	CONSENSUS	High Resolution	0.01	0.7912	2962562
Cohort2012	CONSENSUS	High Resolution	0.05	0.8382	9267817
Cohort2012	CONSENSUS	High Resolution	0.1	0.8878	8469696
Cohort2012	CONSENSUS	High Resolution	0.2	0.9215	129077490
Cohort2012	CONSENSUS	High Resolution	0.5	0.9465	23504726
INTERSECTION	BEAGLE5	phase-states=560	0.0001	0.1305	86071
INTERSECTION	BEAGLE5	phase-states=560	0.0005	0.2633	1181521
INTERSECTION	BEAGLE5	phase-states=560	0.001	0.3826	1186016
INTERSECTION	BEAGLE5	phase-states=560	0.005	0.5139	5460721
INTERSECTION	BEAGLE5	phase-states=560	0.01	0.6029	2962562
INTERSECTION	BEAGLE5	phase-states=560	0.05	0.6761	9267817
INTERSECTION	BEAGLE5	phase-states=560	0.1	0.7345	8469696
INTERSECTION	BEAGLE5	phase-states=560	0.2	0.7908	129077490
INTERSECTION	BEAGLE5	phase-states=560	0.5	0.8428	23504726
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.0001	0.1344	86071

INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.0005	0.2726	1181521
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.001	0.3905	1186016
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.005	0.5187	5460721
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.01	0.6049	2962562
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.05	0.6742	9267817
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.1	0.7313	8469696
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.2	0.7873	129077490
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=8	0.5	0.84	23504726
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.0001	0.1161	86071
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.0005	0.2372	1181521
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.001	0.3514	1186016
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.005	0.4831	5460721
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.01	0.5809	2962562
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.05	0.6646	9267817
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.1	0.7264	8469696
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.2	0.7849	129077490
INTERSECTION	EAGLE2.4.1	Kpbwt=20000	0.5	0.839	23504726
INTERSECTION	CONSENSUS	High Resolution	0.0001	0.1342	86071
INTERSECTION	CONSENSUS	High Resolution	0.0005	0.2722	1181521
INTERSECTION	CONSENSUS	High Resolution	0.001	0.3942	1186016
INTERSECTION	CONSENSUS	High Resolution	0.005	0.5248	5460721
INTERSECTION	CONSENSUS	High Resolution	0.01	0.6153	2962562
INTERSECTION	CONSENSUS	High Resolution	0.05	0.6893	9267817
INTERSECTION	CONSENSUS	High Resolution	0.1	0.7475	8469696
INTERSECTION	CONSENSUS	High Resolution	0.2	0.7994	129077490
INTERSECTION	CONSENSUS	High Resolution	0.5	0.8487	23504726
UNION	BEAGLE5	phase-states=560	0.0001	0.2038	86071
UNION	BEAGLE5	phase-states=560	0.0005	0.3209	1181521
UNION	BEAGLE5	phase-states=560	0.001	0.5786	1186016
UNION	BEAGLE5	phase-states=560	0.005	0.7039	5460721
UNION	BEAGLE5	phase-states=560	0.01	0.7834	2962562
UNION	BEAGLE5	phase-states=560	0.05	0.832	9267817
UNION	BEAGLE5	phase-states=560	0.1	0.8828	8469696
UNION	BEAGLE5	phase-states=560	0.2	0.9181	129077490
UNION	BEAGLE5	phase-states=560	0.5	0.9445	23504726
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.0001	0.1291	86071
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.0005	0.3092	1181521
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.001	0.4345	1186016
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.005	0.4936	5460721
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.01	0.4804	2962562
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.05	0.4388	9267817
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.1	0.4615	8469696
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.2	0.5205	129077490
UNION	SHAPEIT4.1.2	pbwt-depth=8	0.5	0.5892	23504726
UNION	EAGLE2.4.1	Kpbwt=20000	0.0001	0.14	86071
UNION	EAGLE2.4.1	Kpbwt=20000	0.0005	0.3402	1181521
UNION	EAGLE2.4.1	Kpbwt=20000	0.001	0.4914	1186016
UNION	EAGLE2.4.1	Kpbwt=20000	0.005	0.6273	5460721

UNION	EAGLE2.4.1	Kpbwt=20000	0.01	0.7226	2962562
UNION	EAGLE2.4.1	Kpbwt=20000	0.05	0.78	9267817
UNION	EAGLE2.4.1	Kpbwt=20000	0.1	0.8364	8469696
UNION	EAGLE2.4.1	Kpbwt=20000	0.2	0.878	129077490
UNION	EAGLE2.4.1	Kpbwt=20000	0.5	0.9125	23504726
UNION	CONSENSUS	High Resolution	0.0001	0.183	86071
UNION	CONSENSUS	High Resolution	0.0005	0.3952	1181521
UNION	CONSENSUS	High Resolution	0.001	0.5519	1186016
UNION	CONSENSUS	High Resolution	0.005	0.6901	5460721
UNION	CONSENSUS	High Resolution	0.01	0.7625	2962562
UNION	CONSENSUS	High Resolution	0.05	0.7755	9267817
UNION	CONSENSUS	High Resolution	0.1	0.8315	8469696
UNION	CONSENSUS	High Resolution	0.2	0.8736	129077490
UNION	CONSENSUS	High Resolution	0.5	0.9094	23504726
TWOSTAGE	BEAGLE5	phase-states=560	0.0001	0.2125	86071
TWOSTAGE	BEAGLE5	phase-states=560	0.0005	0.4361	1181521
TWOSTAGE	BEAGLE5	phase-states=560	0.001	0.5927	1186016
TWOSTAGE	BEAGLE5	phase-states=560	0.005	0.7194	5460721
TWOSTAGE	BEAGLE5	phase-states=560	0.01	0.8012	2962562
TWOSTAGE	BEAGLE5	phase-states=560	0.05	0.8555	9267817
TWOSTAGE	BEAGLE5	phase-states=560	0.1	0.8966	8469696
TWOSTAGE	BEAGLE5	phase-states=560	0.2	0.9249	129077490
TWOSTAGE	BEAGLE5	phase-states=560	0.5	0.9468	23504726
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.0001	0.2148	86071
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.0005	0.4392	1181521
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.001	0.5952	1186016
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.005	0.7206	5460721
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.01	0.8025	2962562
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.05	0.8561	9267817
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.1	0.897	8469696
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.2	0.9252	129077490
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=8	0.5	0.9469	23504726
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.0001	0.2077	86071
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.0005	0.4314	1181521
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.001	0.5887	1186016
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.005	0.7158	5460721
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.01	0.7991	2962562
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.05	0.8542	9267817
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.1	0.8956	8469696
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.2	0.9244	129077490
TWOSTAGE	EAGLE2.4.1	Kpbwt=20000	0.5	0.9464	23504726
TWOSTAGE	CONSENSUS	High Resolution	0.0001	0.2164	86071
TWOSTAGE	CONSENSUS	High Resolution	0.0005	0.4397	1181521
TWOSTAGE	CONSENSUS	High Resolution	0.001	0.5963	1186016
TWOSTAGE	CONSENSUS	High Resolution	0.005	0.7216	5460721
TWOSTAGE	CONSENSUS	High Resolution	0.01	0.8037	2962562
TWOSTAGE	CONSENSUS	High Resolution	0.05	0.8571	9267817
TWOSTAGE	CONSENSUS	High Resolution	0.1	0.898	8469696

TWOSTAGE	CONSENSUS	High Resolution	0.2	0.926	129077490
TWOSTAGE	CONSENSUS	High Resolution	0.5	0.9474	23504726
Cohort2015i	BEAGLE5	phase-states=560	0.0001	0.2288	86071
Cohort2015i	BEAGLE5	phase-states=560	0.0005	0.4653	1181521
Cohort2015i	BEAGLE5	phase-states=560	0.001	0.6241	1186016
Cohort2015i	BEAGLE5	phase-states=560	0.005	0.7503	5460721
Cohort2015i	BEAGLE5	phase-states=560	0.01	0.8212	2962562
Cohort2015i	BEAGLE5	phase-states=560	0.05	0.8931	9267817
Cohort2015i	BEAGLE5	phase-states=560	0.1	0.9142	8469696
Cohort2015i	BEAGLE5	phase-states=560	0.2	0.9308	129077490
Cohort2015i	BEAGLE5	phase-states=560	0.5	0.9465	23504726
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.0001	0.2358	86071
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.0005	0.4712	1181521
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.001	0.6308	1186016
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.005	0.7516	5460721
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.01	0.8187	2962562
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.05	0.893	9267817
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.1	0.914	8469696
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.2	0.9303	129077490
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=8	0.5	0.9464	23504726
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.0001	0.2262	86071
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.0005	0.4535	1181521
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.001	0.6147	1186016
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.005	0.74	5460721
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.01	0.8129	2962562
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.05	0.8904	9267817
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.1	0.912	8469696
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.2	0.9288	129077490
Cohort2015i	EAGLE2.4.1	Kpbwt=20000	0.5	0.9452	23504726
Cohort2015i	CONSENSUS	High Resolution	0.0001	0.2345	86071
Cohort2015i	CONSENSUS	High Resolution	0.0005	0.4737	1181521
Cohort2015i	CONSENSUS	High Resolution	0.001	0.6351	1186016
Cohort2015i	CONSENSUS	High Resolution	0.005	0.7582	5460721
Cohort2015i	CONSENSUS	High Resolution	0.01	0.8263	2962562
Cohort2015i	CONSENSUS	High Resolution	0.05	0.8967	9267817
Cohort2015i	CONSENSUS	High Resolution	0.1	0.9179	8469696
Cohort2015i	CONSENSUS	High Resolution	0.2	0.9331	129077490
Cohort2015i	CONSENSUS	High Resolution	0.5	0.9484	23504726
Cohort2015i	BEAGLE5	phase-states=280	0.0001	0.2038	86071
Cohort2015i	BEAGLE5	phase-states=280	0.0005	0.4168	1181521
Cohort2015i	BEAGLE5	phase-states=280	0.001	0.5758	1186016
Cohort2015i	BEAGLE5	phase-states=280	0.005	0.7015	5460721
Cohort2015i	BEAGLE5	phase-states=280	0.01	0.7799	2962562
Cohort2015i	BEAGLE5	phase-states=280	0.05	0.8308	9267817
Cohort2015i	BEAGLE5	phase-states=280	0.1	0.8817	8469696
Cohort2015i	BEAGLE5	phase-states=280	0.2	0.9173	129077490
Cohort2015i	BEAGLE5	phase-states=280	0.5	0.9444	23504726
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.0001	0.207	86071

Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.0005	0.4244	1181521
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.001	0.5779	1186016
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.005	0.7042	5460721
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.01	0.7826	2962562
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.05	0.8285	9267817
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.1	0.8796	8469696
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.2	0.9159	129077490
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.5	0.9432	23504726
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.0001	0.1891	86071
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.0005	0.3997	1181521
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.001	0.5531	1186016
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.005	0.6853	5460721
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.01	0.7708	2962562
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.05	0.8236	9267817
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.1	0.8774	8469696
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.2	0.9142	129077490
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.5	0.9423	23504726
Cohort2015i	CONSENSUS	Default	0.0001	0.209	86071
Cohort2015i	CONSENSUS	Default	0.0005	0.4306	1181521
Cohort2015i	CONSENSUS	Default	0.001	0.5681	1186016
Cohort2015i	CONSENSUS	Default	0.005	0.7128	5460721
Cohort2015i	CONSENSUS	Default	0.01	0.7915	2962562
Cohort2015i	CONSENSUS	Default	0.05	0.8384	9267817
Cohort2015i	CONSENSUS	Default	0.1	0.8875	8469696
Cohort2015i	CONSENSUS	Default	0.2	0.9212	129077490
Cohort2015i	CONSENSUS	Default	0.5	0.9465	23504726
INTERSECTION	BEAGLE5	phase-states=280	0.0001	0.1276	86071
INTERSECTION	BEAGLE5	phase-states=280	0.0005	0.2566	1181521
INTERSECTION	BEAGLE5	phase-states=280	0.001	0.3734	1186016
INTERSECTION	BEAGLE5	phase-states=280	0.005	0.5105	5460721
INTERSECTION	BEAGLE5	phase-states=280	0.01	0.6023	2962562
INTERSECTION	BEAGLE5	phase-states=280	0.05	0.674	9267817
INTERSECTION	BEAGLE5	phase-states=280	0.1	0.7317	8469696
INTERSECTION	BEAGLE5	phase-states=280	0.2	0.7879	129077490
INTERSECTION	BEAGLE5	phase-states=280	0.5	0.841	23504726
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.0001	0.1246	86071
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.0005	0.2578	1181521
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.001	0.3781	1186016
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.005	0.5059	5460721
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.01	0.5943	2962562
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.05	0.6635	9267817
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.1	0.724	8469696
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.2	0.7825	129077490
INTERSECTION	SHAPEIT4.1.2	pbwt-depth=4	0.5	0.8359	23504726
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.0001	0.1145	86071
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.0005	0.238	1181521
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.001	0.3596	1186016
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.005	0.4902	5460721

INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.01	0.5845	2962562
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.05	0.6642	9267817
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.1	0.7261	8469696
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.2	0.7834	129077490
INTERSECTION	EAGLE2.4.1	Kpbwt=10000	0.5	0.8377	23504726
INTERSECTION	CONSENSUS	Default	0.0001	0.1327	86071
INTERSECTION	CONSENSUS	Default	0.0005	0.2668	1181521
INTERSECTION	CONSENSUS	Default	0.001	0.3907	1186016
INTERSECTION	CONSENSUS	Default	0.005	0.5208	5460721
INTERSECTION	CONSENSUS	Default	0.01	0.612	2962562
INTERSECTION	CONSENSUS	Default	0.05	0.685	9267817
INTERSECTION	CONSENSUS	Default	0.1	0.7431	8469696
INTERSECTION	CONSENSUS	Default	0.2	0.7976	129077490
INTERSECTION	CONSENSUS	Default	0.5	0.8474	23504726
UNION	BEAGLE5	phase-states=280	0.0001	0.1793	86071
UNION	BEAGLE5	phase-states=280	0.0005	0.3825	1181521
UNION	BEAGLE5	phase-states=280	0.001	0.5464	1186016
UNION	BEAGLE5	phase-states=280	0.005	0.6832	5460721
UNION	BEAGLE5	phase-states=280	0.01	0.7609	2962562
UNION	BEAGLE5	phase-states=280	0.05	0.8164	9267817
UNION	BEAGLE5	phase-states=280	0.1	0.865	8469696
UNION	BEAGLE5	phase-states=280	0.2	0.9017	129077490
UNION	BEAGLE5	phase-states=280	0.5	0.9305	23504726
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.0001	0.1147	86071
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.0005	0.298	1181521
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.001	0.4192	1186016
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.005	0.4712	5460721
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.01	0.4552	2962562
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.05	0.4088	9267817
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.1	0.4334	8469696
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.2	0.4902	129077490
UNION	SHAPEIT4.1.2	pbwt-depth=4	0.5	0.5597	23504726
UNION	EAGLE2.4.1	Kpbwt=10000	0.0001	0.1387	86071
UNION	EAGLE2.4.1	Kpbwt=10000	0.0005	0.3352	1181521
UNION	EAGLE2.4.1	Kpbwt=10000	0.001	0.4895	1186016
UNION	EAGLE2.4.1	Kpbwt=10000	0.005	0.629	5460721
UNION	EAGLE2.4.1	Kpbwt=10000	0.01	0.7212	2962562
UNION	EAGLE2.4.1	Kpbwt=10000	0.05	0.7755	9267817
UNION	EAGLE2.4.1	Kpbwt=10000	0.1	0.8315	8469696
UNION	EAGLE2.4.1	Kpbwt=10000	0.2	0.8736	129077490
UNION	EAGLE2.4.1	Kpbwt=10000	0.5	0.9094	23504726
UNION	CONSENSUS	Default	0.0001	0.1703	86071
UNION	CONSENSUS	Default	0.0005	0.3782	1181521
UNION	CONSENSUS	Default	0.001	0.542	1186016
UNION	CONSENSUS	Default	0.005	0.6771	5460721
UNION	CONSENSUS	Default	0.01	0.7522	2962562
UNION	CONSENSUS	Default	0.05	0.7986	9267817
UNION	CONSENSUS	Default	0.1	0.8495	8469696

UNION	CONSENSUS	Default	0.2	0.8876	129077490
UNION	CONSENSUS	Default	0.5	0.9202	23504726
TWOSTAGE	BEAGLE5	phase-states=280	0.0001	0.2119	86071
TWOSTAGE	BEAGLE5	phase-states=280	0.0005	0.4347	1181521
TWOSTAGE	BEAGLE5	phase-states=280	0.001	0.59	1186016
TWOSTAGE	BEAGLE5	phase-states=280	0.005	0.7191	5460721
TWOSTAGE	BEAGLE5	phase-states=280	0.01	0.7995	2962562
TWOSTAGE	BEAGLE5	phase-states=280	0.05	0.8554	9267817
TWOSTAGE	BEAGLE5	phase-states=280	0.1	0.8967	8469696
TWOSTAGE	BEAGLE5	phase-states=280	0.2	0.9251	129077490
TWOSTAGE	BEAGLE5	phase-states=280	0.5	0.9468	23504726
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.0001	0.2145	86071
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.0005	0.4375	1181521
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.001	0.5928	1186016
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.005	0.7186	5460721
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.01	0.8003	2962562
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.05	0.8553	9267817
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.1	0.8962	8469696
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.2	0.9249	129077490
TWOSTAGE	SHAPEIT4.1.2	pbwt-depth=4	0.5	0.9466	23504726
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.0001	0.2066	86071
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.0005	0.4297	1181521
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.001	0.588	1186016
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.005	0.7152	5460721
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.01	0.7984	2962562
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.05	0.8533	9267817
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.1	0.8948	8469696
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.2	0.9242	129077490
TWOSTAGE	EAGLE2.4.1	Kpbwt=10000	0.5	0.9461	23504726
TWOSTAGE	CONSENSUS	Default	0.0001	0.216	86071
TWOSTAGE	CONSENSUS	Default	0.0005	0.4396	1181521
TWOSTAGE	CONSENSUS	Default	0.001	0.5963	1186016
TWOSTAGE	CONSENSUS	Default	0.005	0.7218	5460721
TWOSTAGE	CONSENSUS	Default	0.01	0.8042	2962562
TWOSTAGE	CONSENSUS	Default	0.05	0.8574	9267817
TWOSTAGE	CONSENSUS	Default	0.1	0.8981	8469696
TWOSTAGE	CONSENSUS	Default	0.2	0.9263	129077490
TWOSTAGE	CONSENSUS	Default	0.5	0.9475	23504726
Cohort2015i	BEAGLE5	phase-states=280	0.0001	0.229	86071
Cohort2015i	BEAGLE5	phase-states=280	0.0005	0.4617	1181521
Cohort2015i	BEAGLE5	phase-states=280	0.001	0.6185	1186016
Cohort2015i	BEAGLE5	phase-states=280	0.005	0.7466	5460721
Cohort2015i	BEAGLE5	phase-states=280	0.01	0.8175	2962562
Cohort2015i	BEAGLE5	phase-states=280	0.05	0.8906	9267817
Cohort2015i	BEAGLE5	phase-states=280	0.1	0.9129	8469696
Cohort2015i	BEAGLE5	phase-states=280	0.2	0.9297	129077490
Cohort2015i	BEAGLE5	phase-states=280	0.5	0.9456	23504726
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.0001	0.2344	86071

Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.0005	0.4681	1181521
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.001	0.6281	1186016
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.005	0.7483	5460721
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.01	0.8165	2962562
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.05	0.8911	9267817
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.1	0.9118	8469696
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.2	0.9291	129077490
Cohort2015i	SHAPEIT4.1.2	pbwt-depth=4	0.5	0.9452	23504726
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.0001	0.2265	86071
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.0005	0.4537	1181521
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.001	0.6164	1186016
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.005	0.7414	5460721
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.01	0.8126	2962562
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.05	0.8893	9267817
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.1	0.9113	8469696
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.2	0.9282	129077490
Cohort2015i	EAGLE2.4.1	Kpbwt=10000	0.5	0.945	23504726
Cohort2015i	CONSENSUS	Default	0.0001	0.2392	86071
Cohort2015i	CONSENSUS	Default	0.0005	0.4747	1181521
Cohort2015i	CONSENSUS	Default	0.001	0.6362	1186016
Cohort2015i	CONSENSUS	Default	0.005	0.7578	5460721
Cohort2015i	CONSENSUS	Default	0.01	0.8259	2962562
Cohort2015i	CONSENSUS	Default	0.05	0.896	9267817
Cohort2015i	CONSENSUS	Default	0.1	0.9171	8469696
Cohort2015i	CONSENSUS	Default	0.2	0.9328	129077490
Cohort2015i	CONSENSUS	Default	0.5	0.9483	23504726

Supplementary Table 10 Switch Error Rates By Chromosome

PROTOCOL	TOOL	PARAMETERS	CHR	SER
COHORT2012	BEAGLE5	PhaseStates560	1	0.3093
COHORT2012	BEAGLE5	PhaseStates560	2	0.2328
COHORT2012	BEAGLE5	PhaseStates560	3	0.2581
COHORT2012	BEAGLE5	PhaseStates560	4	0.2275
COHORT2012	BEAGLE5	PhaseStates560	5	0.2778
COHORT2012	BEAGLE5	PhaseStates560	6	0.2012
COHORT2012	BEAGLE5	PhaseStates560	7	0.2293
COHORT2012	BEAGLE5	PhaseStates560	8	0.2432
COHORT2012	BEAGLE5	PhaseStates560	9	0.3757
COHORT2012	BEAGLE5	PhaseStates560	10	0.2758
COHORT2012	BEAGLE5	PhaseStates560	11	0.28
COHORT2012	BEAGLE5	PhaseStates560	12	0.2673
COHORT2012	BEAGLE5	PhaseStates560	13	0.2485
COHORT2012	BEAGLE5	PhaseStates560	14	0.2814
COHORT2012	BEAGLE5	PhaseStates560	15	0.4269
COHORT2012	BEAGLE5	PhaseStates560	16	0.3655
COHORT2012	BEAGLE5	PhaseStates560	17	0.4982
COHORT2012	BEAGLE5	PhaseStates560	18	0.3489
COHORT2012	BEAGLE5	PhaseStates560	19	0.7299
COHORT2012	BEAGLE5	PhaseStates560	20	0.3986
COHORT2012	BEAGLE5	PhaseStates560	21	0.4632
COHORT2012	BEAGLE5	PhaseStates560	22	0.632
COHORT2015i	BEAGLE5	PhaseStates560	1	0.2652
COHORT2015i	BEAGLE5	PhaseStates560	2	0.2631
COHORT2015i	BEAGLE5	PhaseStates560	3	0.2553
COHORT2015i	BEAGLE5	PhaseStates560	4	0.2792
COHORT2015i	BEAGLE5	PhaseStates560	5	0.2625
COHORT2015i	BEAGLE5	PhaseStates560	6	0.1917
COHORT2015i	BEAGLE5	PhaseStates560	7	0.2459
COHORT2015i	BEAGLE5	PhaseStates560	8	0.2718
COHORT2015i	BEAGLE5	PhaseStates560	9	0.3644
COHORT2015i	BEAGLE5	PhaseStates560	10	0.2666
COHORT2015i	BEAGLE5	PhaseStates560	11	0.2388
COHORT2015i	BEAGLE5	PhaseStates560	12	0.2459
COHORT2015i	BEAGLE5	PhaseStates560	13	0.2665
COHORT2015i	BEAGLE5	PhaseStates560	14	0.3257
COHORT2015i	BEAGLE5	PhaseStates560	15	0.4697
COHORT2015i	BEAGLE5	PhaseStates560	16	0.4223
COHORT2015i	BEAGLE5	PhaseStates560	17	0.4995
COHORT2015i	BEAGLE5	PhaseStates560	18	0.3334
COHORT2015i	BEAGLE5	PhaseStates560	19	0.9439
COHORT2015i	BEAGLE5	PhaseStates560	20	0.4649
COHORT2015i	BEAGLE5	PhaseStates560	21	0.3993
COHORT2015i	BEAGLE5	PhaseStates560	22	0.6673
COHORT2012	BEAGLE5	PhaseStates280	1	0.2933
COHORT2012	BEAGLE5	PhaseStates280	2	0.2576

COHORT2012	BEAGLE5	PhaseStates280	3	0.264
COHORT2012	BEAGLE5	PhaseStates280	4	0.2253
COHORT2012	BEAGLE5	PhaseStates280	5	0.2987
COHORT2012	BEAGLE5	PhaseStates280	6	0.2193
COHORT2012	BEAGLE5	PhaseStates280	7	0.2466
COHORT2012	BEAGLE5	PhaseStates280	8	0.2398
COHORT2012	BEAGLE5	PhaseStates280	9	0.4037
COHORT2012	BEAGLE5	PhaseStates280	10	0.2758
COHORT2012	BEAGLE5	PhaseStates280	11	0.2815
COHORT2012	BEAGLE5	PhaseStates280	12	0.2631
COHORT2012	BEAGLE5	PhaseStates280	13	0.2946
COHORT2012	BEAGLE5	PhaseStates280	14	0.2992
COHORT2012	BEAGLE5	PhaseStates280	15	0.409
COHORT2012	BEAGLE5	PhaseStates280	16	0.3392
COHORT2012	BEAGLE5	PhaseStates280	17	0.5083
COHORT2012	BEAGLE5	PhaseStates280	18	0.3462
COHORT2012	BEAGLE5	PhaseStates280	19	0.7547
COHORT2012	BEAGLE5	PhaseStates280	20	0.4358
COHORT2012	BEAGLE5	PhaseStates280	21	0.4408
COHORT2012	BEAGLE5	PhaseStates280	22	0.7296
COHORT2015i	BEAGLE5	PhaseStates280	1	0.2804
COHORT2015i	BEAGLE5	PhaseStates280	2	0.2575
COHORT2015i	BEAGLE5	PhaseStates280	3	0.262
COHORT2015i	BEAGLE5	PhaseStates280	4	0.2739
COHORT2015i	BEAGLE5	PhaseStates280	5	0.2767
COHORT2015i	BEAGLE5	PhaseStates280	6	0.1913
COHORT2015i	BEAGLE5	PhaseStates280	7	0.2487
COHORT2015i	BEAGLE5	PhaseStates280	8	0.2738
COHORT2015i	BEAGLE5	PhaseStates280	9	0.3679
COHORT2015i	BEAGLE5	PhaseStates280	10	0.2784
COHORT2015i	BEAGLE5	PhaseStates280	11	0.2388
COHORT2015i	BEAGLE5	PhaseStates280	12	0.2491
COHORT2015i	BEAGLE5	PhaseStates280	13	0.2919
COHORT2015i	BEAGLE5	PhaseStates280	14	0.3217
COHORT2015i	BEAGLE5	PhaseStates280	15	0.4486
COHORT2015i	BEAGLE5	PhaseStates280	16	0.4378
COHORT2015i	BEAGLE5	PhaseStates280	17	0.5075
COHORT2015i	BEAGLE5	PhaseStates280	18	0.3973
COHORT2015i	BEAGLE5	PhaseStates280	19	0.9221
COHORT2015i	BEAGLE5	PhaseStates280	20	0.5009
COHORT2015i	BEAGLE5	PhaseStates280	21	0.4455
COHORT2015i	BEAGLE5	PhaseStates280	22	0.6739
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	1	0.272
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	2	0.2599
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	3	0.2411
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	4	0.2208
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	5	0.2603
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	6	0.1949

COHORT2012	SHAPEIT4.1.2	PbwtDepth8	7	0.2312
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	8	0.2239
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	9	0.3556
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	10	0.246
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	11	0.2646
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	12	0.2694
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	13	0.262
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	14	0.2693
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	15	0.4064
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	16	0.368
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	17	0.5103
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	18	0.4049
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	19	0.797
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	20	0.41
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	21	0.452
COHORT2012	SHAPEIT4.1.2	PbwtDepth8	22	0.6298
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	1	0.2417
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	2	0.2126
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	3	0.2155
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	4	0.2295
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	5	0.2314
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	6	0.1832
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	7	0.2086
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	8	0.1908
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	9	0.3338
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	10	0.2533
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	11	0.2174
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	12	0.2272
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	13	0.2397
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	14	0.2633
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	15	0.4131
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	16	0.3644
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	17	0.4754
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	18	0.3406
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	19	0.9178
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	20	0.4516
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	21	0.3882
COHORT2015i	SHAPEIT4.1.2	PbwtDepth8	22	0.6629
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	1	0.3232
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	2	0.2599
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	3	0.2632
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	4	0.2471
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	5	0.2952
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	6	0.2038
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	7	0.2629
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	8	0.2536
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	9	0.387
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	10	0.2748

COHORT2012	SHAPEIT4.1.2	PbwtDepth4	11	0.2743
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	12	0.3317
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	13	0.3244
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	14	0.2959
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	15	0.4838
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	16	0.4123
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	17	0.583
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	18	0.4465
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	19	0.9197
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	20	0.4926
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	21	0.4913
COHORT2012	SHAPEIT4.1.2	PbwtDepth4	22	0.8206
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	1	0.2889
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	2	0.2447
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	3	0.2502
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	4	0.2668
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	5	0.2527
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	6	0.1792
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	7	0.2239
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	8	0.2225
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	9	0.3885
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	10	0.2631
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	11	0.2449
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	12	0.2475
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	13	0.2835
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	14	0.3001
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	15	0.4486
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	16	0.4198
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	17	0.5697
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	18	0.3541
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	19	1.2004
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	20	0.4824
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	21	0.4843
COHORT2015i	SHAPEIT4.1.2	PbwtDepth4	22	0.8693
COHORT2012	EAGLE2.4.1	Kpbwt20000	1	0.5431
COHORT2012	EAGLE2.4.1	Kpbwt20000	2	0.4732
COHORT2012	EAGLE2.4.1	Kpbwt20000	3	0.4953
COHORT2012	EAGLE2.4.1	Kpbwt20000	4	0.4603
COHORT2012	EAGLE2.4.1	Kpbwt20000	5	0.5193
COHORT2012	EAGLE2.4.1	Kpbwt20000	6	0.5895
COHORT2012	EAGLE2.4.1	Kpbwt20000	7	0.4576
COHORT2012	EAGLE2.4.1	Kpbwt20000	8	0.4123
COHORT2012	EAGLE2.4.1	Kpbwt20000	9	0.6435
COHORT2012	EAGLE2.4.1	Kpbwt20000	10	0.4889
COHORT2012	EAGLE2.4.1	Kpbwt20000	11	0.4704
COHORT2012	EAGLE2.4.1	Kpbwt20000	12	0.5047
COHORT2012	EAGLE2.4.1	Kpbwt20000	13	0.4373
COHORT2012	EAGLE2.4.1	Kpbwt20000	14	0.4605

COHORT2012	EAGLE2.4.1	Kpbwt20000	15	0.7687
COHORT2012	EAGLE2.4.1	Kpbwt20000	16	0.6169
COHORT2012	EAGLE2.4.1	Kpbwt20000	17	0.8563
COHORT2012	EAGLE2.4.1	Kpbwt20000	18	0.5604
COHORT2012	EAGLE2.4.1	Kpbwt20000	19	1.2335
COHORT2012	EAGLE2.4.1	Kpbwt20000	20	0.6991
COHORT2012	EAGLE2.4.1	Kpbwt20000	21	0.6258
COHORT2012	EAGLE2.4.1	Kpbwt20000	22	1.0068
COHORT2015i	EAGLE2.4.1	Kpbwt20000	1	0.5023
COHORT2015i	EAGLE2.4.1	Kpbwt20000	2	0.6128
COHORT2015i	EAGLE2.4.1	Kpbwt20000	3	0.6071
COHORT2015i	EAGLE2.4.1	Kpbwt20000	4	0.7147
COHORT2015i	EAGLE2.4.1	Kpbwt20000	5	0.6832
COHORT2015i	EAGLE2.4.1	Kpbwt20000	6	0.8104
COHORT2015i	EAGLE2.4.1	Kpbwt20000	7	0.6455
COHORT2015i	EAGLE2.4.1	Kpbwt20000	8	0.5929
COHORT2015i	EAGLE2.4.1	Kpbwt20000	9	0.9234
COHORT2015i	EAGLE2.4.1	Kpbwt20000	10	0.6758
COHORT2015i	EAGLE2.4.1	Kpbwt20000	11	0.5597
COHORT2015i	EAGLE2.4.1	Kpbwt20000	12	0.65
COHORT2015i	EAGLE2.4.1	Kpbwt20000	13	0.7108
COHORT2015i	EAGLE2.4.1	Kpbwt20000	14	0.761
COHORT2015i	EAGLE2.4.1	Kpbwt20000	15	0.964
COHORT2015i	EAGLE2.4.1	Kpbwt20000	16	1.0223
COHORT2015i	EAGLE2.4.1	Kpbwt20000	17	1.331
COHORT2015i	EAGLE2.4.1	Kpbwt20000	18	0.7909
COHORT2015i	EAGLE2.4.1	Kpbwt20000	19	2.1985
COHORT2015i	EAGLE2.4.1	Kpbwt20000	20	1.1377
COHORT2015i	EAGLE2.4.1	Kpbwt20000	21	1.0186
COHORT2015i	EAGLE2.4.1	Kpbwt20000	22	1.6994
COHORT2012	EAGLE2.4.1	Kpbwt10000	1	0.6478
COHORT2012	EAGLE2.4.1	Kpbwt10000	2	0.5603
COHORT2012	EAGLE2.4.1	Kpbwt10000	3	0.5884
COHORT2012	EAGLE2.4.1	Kpbwt10000	4	0.5491
COHORT2012	EAGLE2.4.1	Kpbwt10000	5	0.606
COHORT2012	EAGLE2.4.1	Kpbwt10000	6	0.8292
COHORT2012	EAGLE2.4.1	Kpbwt10000	7	0.4956
COHORT2012	EAGLE2.4.1	Kpbwt10000	8	0.4682
COHORT2012	EAGLE2.4.1	Kpbwt10000	9	0.6827
COHORT2012	EAGLE2.4.1	Kpbwt10000	10	0.5557
COHORT2012	EAGLE2.4.1	Kpbwt10000	11	0.5096
COHORT2012	EAGLE2.4.1	Kpbwt10000	12	0.5754
COHORT2012	EAGLE2.4.1	Kpbwt10000	13	0.4827
COHORT2012	EAGLE2.4.1	Kpbwt10000	14	0.5096
COHORT2012	EAGLE2.4.1	Kpbwt10000	15	0.8299
COHORT2012	EAGLE2.4.1	Kpbwt10000	16	0.6621
COHORT2012	EAGLE2.4.1	Kpbwt10000	17	0.9773
COHORT2012	EAGLE2.4.1	Kpbwt10000	18	0.6309

COHORT2012	EAGLE2.4.1	Kpbwt10000	19	1.3561
COHORT2012	EAGLE2.4.1	Kpbwt10000	20	0.789
COHORT2012	EAGLE2.4.1	Kpbwt10000	21	0.6295
COHORT2012	EAGLE2.4.1	Kpbwt10000	22	0.9891
COHORT2015i	EAGLE2.4.1	Kpbwt10000	1	0.6195
COHORT2015i	EAGLE2.4.1	Kpbwt10000	2	0.7837
COHORT2015i	EAGLE2.4.1	Kpbwt10000	3	0.7903
COHORT2015i	EAGLE2.4.1	Kpbwt10000	4	0.8874
COHORT2015i	EAGLE2.4.1	Kpbwt10000	5	0.8649
COHORT2015i	EAGLE2.4.1	Kpbwt10000	6	1.2234
COHORT2015i	EAGLE2.4.1	Kpbwt10000	7	0.7724
COHORT2015i	EAGLE2.4.1	Kpbwt10000	8	0.7398
COHORT2015i	EAGLE2.4.1	Kpbwt10000	9	1.0968
COHORT2015i	EAGLE2.4.1	Kpbwt10000	10	0.7866
COHORT2015i	EAGLE2.4.1	Kpbwt10000	11	0.6658
COHORT2015i	EAGLE2.4.1	Kpbwt10000	12	0.8627
COHORT2015i	EAGLE2.4.1	Kpbwt10000	13	0.8271
COHORT2015i	EAGLE2.4.1	Kpbwt10000	14	0.9315
COHORT2015i	EAGLE2.4.1	Kpbwt10000	15	1.171
COHORT2015i	EAGLE2.4.1	Kpbwt10000	16	1.1691
COHORT2015i	EAGLE2.4.1	Kpbwt10000	17	1.5407
COHORT2015i	EAGLE2.4.1	Kpbwt10000	18	1.0066
COHORT2015i	EAGLE2.4.1	Kpbwt10000	19	2.4387
COHORT2015i	EAGLE2.4.1	Kpbwt10000	20	1.2961
COHORT2015i	EAGLE2.4.1	Kpbwt10000	21	1.1997
COHORT2015i	EAGLE2.4.1	Kpbwt10000	22	1.9366
COHORT2012	CONSENSUS	High Resolution	1	0.2581
COHORT2012	CONSENSUS	High Resolution	2	0.2046
COHORT2012	CONSENSUS	High Resolution	3	0.2151
COHORT2012	CONSENSUS	High Resolution	4	0.1793
COHORT2012	CONSENSUS	High Resolution	5	0.2335
COHORT2012	CONSENSUS	High Resolution	6	0.1775
COHORT2012	CONSENSUS	High Resolution	7	0.2014
COHORT2012	CONSENSUS	High Resolution	8	0.2086
COHORT2012	CONSENSUS	High Resolution	9	0.3122
COHORT2012	CONSENSUS	High Resolution	10	0.2265
COHORT2012	CONSENSUS	High Resolution	11	0.2259
COHORT2012	CONSENSUS	High Resolution	12	0.2259
COHORT2012	CONSENSUS	High Resolution	13	0.2016
COHORT2012	CONSENSUS	High Resolution	14	0.2234
COHORT2012	CONSENSUS	High Resolution	15	0.3597
COHORT2012	CONSENSUS	High Resolution	16	0.2129
COHORT2012	CONSENSUS	High Resolution	17	0.4357
COHORT2012	CONSENSUS	High Resolution	18	0.3073
COHORT2012	CONSENSUS	High Resolution	19	0.689
COHORT2012	CONSENSUS	High Resolution	20	0.3356
COHORT2012	CONSENSUS	High Resolution	21	0.3754
COHORT2012	CONSENSUS	High Resolution	22	0.55

COHORT2015i	CONSENSUS	High Resolution	1	0.2143
COHORT2015i	CONSENSUS	High Resolution	2	0.2093
COHORT2015i	CONSENSUS	High Resolution	3	0.2076
COHORT2015i	CONSENSUS	High Resolution	4	0.2313
COHORT2015i	CONSENSUS	High Resolution	5	0.215
COHORT2015i	CONSENSUS	High Resolution	6	0.1788
COHORT2015i	CONSENSUS	High Resolution	7	0.2034
COHORT2015i	CONSENSUS	High Resolution	8	0.1957
COHORT2015i	CONSENSUS	High Resolution	9	0.3303
COHORT2015i	CONSENSUS	High Resolution	10	0.2176
COHORT2015i	CONSENSUS	High Resolution	11	0.2071
COHORT2015i	CONSENSUS	High Resolution	12	0.1997
COHORT2015i	CONSENSUS	High Resolution	13	0.227
COHORT2015i	CONSENSUS	High Resolution	14	0.2593
COHORT2015i	CONSENSUS	High Resolution	15	0.3692
COHORT2015i	CONSENSUS	High Resolution	16	0.3611
COHORT2015i	CONSENSUS	High Resolution	17	0.4554
COHORT2015i	CONSENSUS	High Resolution	18	0.3101
COHORT2015i	CONSENSUS	High Resolution	19	0.8536
COHORT2015i	CONSENSUS	High Resolution	20	0.4043
COHORT2015i	CONSENSUS	High Resolution	21	0.3623
COHORT2015i	CONSENSUS	High Resolution	22	0.6146
COHORT2012	CONSENSUS	Default	1	0.2517
COHORT2012	CONSENSUS	Default	2	0.2146
COHORT2012	CONSENSUS	Default	3	0.2218
COHORT2012	CONSENSUS	Default	4	0.1998
COHORT2012	CONSENSUS	Default	5	0.2411
COHORT2012	CONSENSUS	Default	6	0.1823
COHORT2012	CONSENSUS	Default	7	0.2033
COHORT2012	CONSENSUS	Default	8	0.215
COHORT2012	CONSENSUS	Default	9	0.3182
COHORT2012	CONSENSUS	Default	10	0.2311
COHORT2012	CONSENSUS	Default	11	0.227
COHORT2012	CONSENSUS	Default	12	0.2453
COHORT2012	CONSENSUS	Default	13	0.2165
COHORT2012	CONSENSUS	Default	14	0.2476
COHORT2012	CONSENSUS	Default	15	0.3741
COHORT2012	CONSENSUS	Default	16	0.2949
COHORT2012	CONSENSUS	Default	17	0.4438
COHORT2012	CONSENSUS	Default	18	0.3245
COHORT2012	CONSENSUS	Default	19	0.7007
COHORT2012	CONSENSUS	Default	20	0.3883
COHORT2012	CONSENSUS	Default	21	0.3754
COHORT2012	CONSENSUS	Default	22	0.601
COHORT2015i	CONSENSUS	Default	1	0.2378
COHORT2015i	CONSENSUS	Default	2	0.2044
COHORT2015i	CONSENSUS	Default	3	0.2096
COHORT2015i	CONSENSUS	Default	4	0.2432

COHORT2015i	CONSENSUS	Default	5	0.2216
COHORT2015i	CONSENSUS	Default	6	0.1652
COHORT2015i	CONSENSUS	Default	7	0.1957
COHORT2015i	CONSENSUS	Default	8	0.2015
COHORT2015i	CONSENSUS	Default	9	0.3415
COHORT2015i	CONSENSUS	Default	10	0.2293
COHORT2015i	CONSENSUS	Default	11	0.2102
COHORT2015i	CONSENSUS	Default	12	0.2127
COHORT2015i	CONSENSUS	Default	13	0.2574
COHORT2015i	CONSENSUS	Default	14	0.2481
COHORT2015i	CONSENSUS	Default	15	0.3692
COHORT2015i	CONSENSUS	Default	16	0.3848
COHORT2015i	CONSENSUS	Default	17	0.4935
COHORT2015i	CONSENSUS	Default	18	0.3325
COHORT2015i	CONSENSUS	Default	19	0.9163
COHORT2015i	CONSENSUS	Default	20	0.4135
COHORT2015i	CONSENSUS	Default	21	0.4104
COHORT2015i	CONSENSUS	Default	22	0.6651
INTERSECTION	BEAGLE5	PhaseStates560	1	0.417
INTERSECTION	BEAGLE5	PhaseStates560	2	0.3385
INTERSECTION	BEAGLE5	PhaseStates560	3	0.3435
INTERSECTION	BEAGLE5	PhaseStates560	4	0.3132
INTERSECTION	BEAGLE5	PhaseStates560	5	0.3355
INTERSECTION	BEAGLE5	PhaseStates560	6	0.2268
INTERSECTION	BEAGLE5	PhaseStates560	7	0.3128
INTERSECTION	BEAGLE5	PhaseStates560	8	0.3329
INTERSECTION	BEAGLE5	PhaseStates560	9	0.7966
INTERSECTION	BEAGLE5	PhaseStates560	10	0.3199
INTERSECTION	BEAGLE5	PhaseStates560	11	0.4067
INTERSECTION	BEAGLE5	PhaseStates560	12	0.3682
INTERSECTION	BEAGLE5	PhaseStates560	13	0.3432
INTERSECTION	BEAGLE5	PhaseStates560	14	0.4004
INTERSECTION	BEAGLE5	PhaseStates560	15	0.5716
INTERSECTION	BEAGLE5	PhaseStates560	16	0.4992
INTERSECTION	BEAGLE5	PhaseStates560	17	1.0028
INTERSECTION	BEAGLE5	PhaseStates560	18	0.5387
INTERSECTION	BEAGLE5	PhaseStates560	19	1.7933
INTERSECTION	BEAGLE5	PhaseStates560	20	0.88
INTERSECTION	BEAGLE5	PhaseStates560	21	0.6656
INTERSECTION	BEAGLE5	PhaseStates560	22	1.3769
INTERSECTION	BEAGLE5	PhaseStates280	1	0.4177
INTERSECTION	BEAGLE5	PhaseStates280	2	0.3551
INTERSECTION	BEAGLE5	PhaseStates280	3	0.3674
INTERSECTION	BEAGLE5	PhaseStates280	4	0.36
INTERSECTION	BEAGLE5	PhaseStates280	5	0.37
INTERSECTION	BEAGLE5	PhaseStates280	6	0.2429
INTERSECTION	BEAGLE5	PhaseStates280	7	0.2995
INTERSECTION	BEAGLE5	PhaseStates280	8	0.3456

INTERSECTION	BEAGLE5	PhaseStates280	9	0.9086
INTERSECTION	BEAGLE5	PhaseStates280	10	0.325
INTERSECTION	BEAGLE5	PhaseStates280	11	0.4057
INTERSECTION	BEAGLE5	PhaseStates280	12	0.376
INTERSECTION	BEAGLE5	PhaseStates280	13	0.3649
INTERSECTION	BEAGLE5	PhaseStates280	14	0.4402
INTERSECTION	BEAGLE5	PhaseStates280	15	0.5993
INTERSECTION	BEAGLE5	PhaseStates280	16	0.595
INTERSECTION	BEAGLE5	PhaseStates280	17	1.1993
INTERSECTION	BEAGLE5	PhaseStates280	18	0.5673
INTERSECTION	BEAGLE5	PhaseStates280	19	2.1317
INTERSECTION	BEAGLE5	PhaseStates280	20	1.053
INTERSECTION	BEAGLE5	PhaseStates280	21	0.7099
INTERSECTION	BEAGLE5	PhaseStates280	22	1.839
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	1	0.3887
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	2	0.3339
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	3	0.3431
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	4	0.3445
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	5	0.3815
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	6	0.2484
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	7	0.3218
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	8	0.3134
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	9	0.7121
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	10	0.3179
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	11	0.4067
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	12	0.334
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	13	0.3783
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	14	0.4116
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	15	0.6607
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	16	0.5463
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	17	1.0138
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	18	0.5861
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	19	1.8816
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	20	1.0008
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	21	0.78
INTERSECTION	SHAPEIT4.1.2	PbwtDepth8	22	1.7559
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	1	0.5321
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	2	0.4235
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	3	0.4707
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	4	0.4686
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	5	0.4708
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	6	0.289
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	7	0.4126
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	8	0.39
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	9	0.9127
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	10	0.395
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	11	0.5165
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	12	0.4863

INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	13	0.4997
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	14	0.5493
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	15	0.8549
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	16	0.7492
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	17	1.6603
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	18	0.7319
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	19	2.7046
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	20	1.3078
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	21	1.1137
INTERSECTION	SHAPEIT4.1.2	PbwtDepth4	22	2.5291
INTERSECTION	EAGLE2.4.1	Kpbwt20000	1	0.9304
INTERSECTION	EAGLE2.4.1	Kpbwt20000	2	0.7064
INTERSECTION	EAGLE2.4.1	Kpbwt20000	3	0.6898
INTERSECTION	EAGLE2.4.1	Kpbwt20000	4	0.8176
INTERSECTION	EAGLE2.4.1	Kpbwt20000	5	0.8254
INTERSECTION	EAGLE2.4.1	Kpbwt20000	6	0.8322
INTERSECTION	EAGLE2.4.1	Kpbwt20000	7	0.6746
INTERSECTION	EAGLE2.4.1	Kpbwt20000	8	0.6006
INTERSECTION	EAGLE2.4.1	Kpbwt20000	9	1.0699
INTERSECTION	EAGLE2.4.1	Kpbwt20000	10	0.6566
INTERSECTION	EAGLE2.4.1	Kpbwt20000	11	0.5923
INTERSECTION	EAGLE2.4.1	Kpbwt20000	12	0.6686
INTERSECTION	EAGLE2.4.1	Kpbwt20000	13	0.6927
INTERSECTION	EAGLE2.4.1	Kpbwt20000	14	0.8
INTERSECTION	EAGLE2.4.1	Kpbwt20000	15	1.2146
INTERSECTION	EAGLE2.4.1	Kpbwt20000	16	1.0641
INTERSECTION	EAGLE2.4.1	Kpbwt20000	17	2.1383
INTERSECTION	EAGLE2.4.1	Kpbwt20000	18	1.0523
INTERSECTION	EAGLE2.4.1	Kpbwt20000	19	3.101
INTERSECTION	EAGLE2.4.1	Kpbwt20000	20	1.5165
INTERSECTION	EAGLE2.4.1	Kpbwt20000	21	1.1432
INTERSECTION	EAGLE2.4.1	Kpbwt20000	22	2.7658
INTERSECTION	EAGLE2.4.1	Kpbwt10000	1	1.034
INTERSECTION	EAGLE2.4.1	Kpbwt10000	2	0.8064
INTERSECTION	EAGLE2.4.1	Kpbwt10000	3	0.8244
INTERSECTION	EAGLE2.4.1	Kpbwt10000	4	0.9285
INTERSECTION	EAGLE2.4.1	Kpbwt10000	5	0.9659
INTERSECTION	EAGLE2.4.1	Kpbwt10000	6	1.1542
INTERSECTION	EAGLE2.4.1	Kpbwt10000	7	0.7511
INTERSECTION	EAGLE2.4.1	Kpbwt10000	8	0.6722
INTERSECTION	EAGLE2.4.1	Kpbwt10000	9	1.2571
INTERSECTION	EAGLE2.4.1	Kpbwt10000	10	0.7703
INTERSECTION	EAGLE2.4.1	Kpbwt10000	11	0.6711
INTERSECTION	EAGLE2.4.1	Kpbwt10000	12	0.7717
INTERSECTION	EAGLE2.4.1	Kpbwt10000	13	0.7699
INTERSECTION	EAGLE2.4.1	Kpbwt10000	14	0.9242
INTERSECTION	EAGLE2.4.1	Kpbwt10000	15	1.3189
INTERSECTION	EAGLE2.4.1	Kpbwt10000	16	1.1104

INTERSECTION	EAGLE2.4.1	Kpbwt10000	17	2.1632
INTERSECTION	EAGLE2.4.1	Kpbwt10000	18	1.1714
INTERSECTION	EAGLE2.4.1	Kpbwt10000	19	3.2701
INTERSECTION	EAGLE2.4.1	Kpbwt10000	20	1.5247
INTERSECTION	EAGLE2.4.1	Kpbwt10000	21	1.1764
INTERSECTION	EAGLE2.4.1	Kpbwt10000	22	2.8008
INTERSECTION	CONSENSUS	High Resolution	1	0.344
INTERSECTION	CONSENSUS	High Resolution	2	0.278
INTERSECTION	CONSENSUS	High Resolution	3	0.2778
INTERSECTION	CONSENSUS	High Resolution	4	0.2602
INTERSECTION	CONSENSUS	High Resolution	5	0.2988
INTERSECTION	CONSENSUS	High Resolution	6	0.1975
INTERSECTION	CONSENSUS	High Resolution	7	0.2648
INTERSECTION	CONSENSUS	High Resolution	8	0.2564
INTERSECTION	CONSENSUS	High Resolution	9	0.6253
INTERSECTION	CONSENSUS	High Resolution	10	0.2651
INTERSECTION	CONSENSUS	High Resolution	11	0.3254
INTERSECTION	CONSENSUS	High Resolution	12	0.2791
INTERSECTION	CONSENSUS	High Resolution	13	0.2765
INTERSECTION	CONSENSUS	High Resolution	14	0.3184
INTERSECTION	CONSENSUS	High Resolution	15	0.4766
INTERSECTION	CONSENSUS	High Resolution	16	0.4075
INTERSECTION	CONSENSUS	High Resolution	17	0.8511
INTERSECTION	CONSENSUS	High Resolution	18	0.4438
INTERSECTION	CONSENSUS	High Resolution	19	1.5648
INTERSECTION	CONSENSUS	High Resolution	20	0.7736
INTERSECTION	CONSENSUS	High Resolution	21	0.5679
INTERSECTION	CONSENSUS	High Resolution	22	1.1844
INTERSECTION	CONSENSUS	Default	1	0.3808
INTERSECTION	CONSENSUS	Default	2	0.3156
INTERSECTION	CONSENSUS	Default	3	0.32
INTERSECTION	CONSENSUS	Default	4	0.3176
INTERSECTION	CONSENSUS	Default	5	0.3505
INTERSECTION	CONSENSUS	Default	6	0.2264
INTERSECTION	CONSENSUS	Default	7	0.2786
INTERSECTION	CONSENSUS	Default	8	0.2964
INTERSECTION	CONSENSUS	Default	9	0.735
INTERSECTION	CONSENSUS	Default	10	0.2777
INTERSECTION	CONSENSUS	Default	11	0.366
INTERSECTION	CONSENSUS	Default	12	0.3257
INTERSECTION	CONSENSUS	Default	13	0.3137
INTERSECTION	CONSENSUS	Default	14	0.3638
INTERSECTION	CONSENSUS	Default	15	0.5371
INTERSECTION	CONSENSUS	Default	16	0.5097
INTERSECTION	CONSENSUS	Default	17	1.0487
INTERSECTION	CONSENSUS	Default	18	0.5073
INTERSECTION	CONSENSUS	Default	19	1.8816
INTERSECTION	CONSENSUS	Default	20	0.8555

INTERSECTION	CONSENSUS	Default	21	0.5956
INTERSECTION	CONSENSUS	Default	22	1.4495
UNION	BEAGLE5	PhaseStates560	1	0.1774
UNION	BEAGLE5	PhaseStates560	2	0.1151
UNION	BEAGLE5	PhaseStates560	3	0.1319
UNION	BEAGLE5	PhaseStates560	4	0.1319
UNION	BEAGLE5	PhaseStates560	5	0.1077
UNION	BEAGLE5	PhaseStates560	6	0.0925
UNION	BEAGLE5	PhaseStates560	7	0.1237
UNION	BEAGLE5	PhaseStates560	8	0.1269
UNION	BEAGLE5	PhaseStates560	9	0.2337
UNION	BEAGLE5	PhaseStates560	10	0.141
UNION	BEAGLE5	PhaseStates560	11	0.1175
UNION	BEAGLE5	PhaseStates560	12	0.1271
UNION	BEAGLE5	PhaseStates560	13	0.1521
UNION	BEAGLE5	PhaseStates560	14	0.1547
UNION	BEAGLE5	PhaseStates560	15	0.2426
UNION	BEAGLE5	PhaseStates560	16	0.1992
UNION	BEAGLE5	PhaseStates560	17	0.3017
UNION	BEAGLE5	PhaseStates560	18	0.2144
UNION	BEAGLE5	PhaseStates560	19	0.7812
UNION	BEAGLE5	PhaseStates560	20	0.2678
UNION	BEAGLE5	PhaseStates560	21	0.2174
UNION	BEAGLE5	PhaseStates560	22	0.5669
UNION	BEAGLE5	PhaseStates280	1	0.1893
UNION	BEAGLE5	PhaseStates280	2	0.119
UNION	BEAGLE5	PhaseStates280	3	0.1453
UNION	BEAGLE5	PhaseStates280	4	0.1407
UNION	BEAGLE5	PhaseStates280	5	0.1306
UNION	BEAGLE5	PhaseStates280	6	0.0977
UNION	BEAGLE5	PhaseStates280	7	0.1389
UNION	BEAGLE5	PhaseStates280	8	0.1162
UNION	BEAGLE5	PhaseStates280	9	0.2349
UNION	BEAGLE5	PhaseStates280	10	0.1553
UNION	BEAGLE5	PhaseStates280	11	0.1256
UNION	BEAGLE5	PhaseStates280	12	0.124
UNION	BEAGLE5	PhaseStates280	13	0.1493
UNION	BEAGLE5	PhaseStates280	14	0.1722
UNION	BEAGLE5	PhaseStates280	15	0.2998
UNION	BEAGLE5	PhaseStates280	16	0.2188
UNION	BEAGLE5	PhaseStates280	17	0.2837
UNION	BEAGLE5	PhaseStates280	18	0.2432
UNION	BEAGLE5	PhaseStates280	19	0.9468
UNION	BEAGLE5	PhaseStates280	20	0.3522
UNION	BEAGLE5	PhaseStates280	21	0.2948
UNION	BEAGLE5	PhaseStates280	22	0.6567
UNION	SHAPEIT4.1.2	PbwtDepth8	1	0.2066
UNION	SHAPEIT4.1.2	PbwtDepth8	2	0.1532

UNION	SHAPEIT4.1.2	PbwtDepth8	3	0.1485
UNION	SHAPEIT4.1.2	PbwtDepth8	4	0.162
UNION	SHAPEIT4.1.2	PbwtDepth8	5	0.1562
UNION	SHAPEIT4.1.2	PbwtDepth8	6	0.1131
UNION	SHAPEIT4.1.2	PbwtDepth8	7	0.1598
UNION	SHAPEIT4.1.2	PbwtDepth8	8	0.1279
UNION	SHAPEIT4.1.2	PbwtDepth8	9	0.2678
UNION	SHAPEIT4.1.2	PbwtDepth8	10	0.2105
UNION	SHAPEIT4.1.2	PbwtDepth8	11	0.1685
UNION	SHAPEIT4.1.2	PbwtDepth8	12	0.1591
UNION	SHAPEIT4.1.2	PbwtDepth8	13	0.2028
UNION	SHAPEIT4.1.2	PbwtDepth8	14	0.2073
UNION	SHAPEIT4.1.2	PbwtDepth8	15	0.3049
UNION	SHAPEIT4.1.2	PbwtDepth8	16	0.2433
UNION	SHAPEIT4.1.2	PbwtDepth8	17	0.3536
UNION	SHAPEIT4.1.2	PbwtDepth8	18	0.2864
UNION	SHAPEIT4.1.2	PbwtDepth8	19	0.9236
UNION	SHAPEIT4.1.2	PbwtDepth8	20	0.3584
UNION	SHAPEIT4.1.2	PbwtDepth8	21	0.2985
UNION	SHAPEIT4.1.2	PbwtDepth8	22	0.7377
UNION	SHAPEIT4.1.2	PbwtDepth4	1	0.233
UNION	SHAPEIT4.1.2	PbwtDepth4	2	0.1525
UNION	SHAPEIT4.1.2	PbwtDepth4	3	0.1666
UNION	SHAPEIT4.1.2	PbwtDepth4	4	0.1885
UNION	SHAPEIT4.1.2	PbwtDepth4	5	0.1739
UNION	SHAPEIT4.1.2	PbwtDepth4	6	0.1307
UNION	SHAPEIT4.1.2	PbwtDepth4	7	0.1817
UNION	SHAPEIT4.1.2	PbwtDepth4	8	0.1748
UNION	SHAPEIT4.1.2	PbwtDepth4	9	0.2608
UNION	SHAPEIT4.1.2	PbwtDepth4	10	0.2115
UNION	SHAPEIT4.1.2	PbwtDepth4	11	0.1757
UNION	SHAPEIT4.1.2	PbwtDepth4	12	0.1426
UNION	SHAPEIT4.1.2	PbwtDepth4	13	0.2
UNION	SHAPEIT4.1.2	PbwtDepth4	14	0.244
UNION	SHAPEIT4.1.2	PbwtDepth4	15	0.3504
UNION	SHAPEIT4.1.2	PbwtDepth4	16	0.2808
UNION	SHAPEIT4.1.2	PbwtDepth4	17	0.3936
UNION	SHAPEIT4.1.2	PbwtDepth4	18	0.3044
UNION	SHAPEIT4.1.2	PbwtDepth4	19	0.1211
UNION	SHAPEIT4.1.2	PbwtDepth4	20	0.4243
UNION	SHAPEIT4.1.2	PbwtDepth4	21	0.339
UNION	SHAPEIT4.1.2	PbwtDepth4	22	0.8953
UNION	EAGLE2.4.1	Kpbwt20000	1	0.6614
UNION	EAGLE2.4.1	Kpbwt20000	2	0.4465
UNION	EAGLE2.4.1	Kpbwt20000	3	0.5062
UNION	EAGLE2.4.1	Kpbwt20000	4	0.5727
UNION	EAGLE2.4.1	Kpbwt20000	5	0.4961
UNION	EAGLE2.4.1	Kpbwt20000	6	0.7248

UNION	EAGLE2.4.1	Kpbwt20000	7	0.4509
UNION	EAGLE2.4.1	Kpbwt20000	8	0.3525
UNION	EAGLE2.4.1	Kpbwt20000	9	0.5908
UNION	EAGLE2.4.1	Kpbwt20000	10	0.47
UNION	EAGLE2.4.1	Kpbwt20000	11	0.3718
UNION	EAGLE2.4.1	Kpbwt20000	12	0.5031
UNION	EAGLE2.4.1	Kpbwt20000	13	0.4337
UNION	EAGLE2.4.1	Kpbwt20000	14	0.5598
UNION	EAGLE2.4.1	Kpbwt20000	15	0.7968
UNION	EAGLE2.4.1	Kpbwt20000	16	0.6123
UNION	EAGLE2.4.1	Kpbwt20000	17	0.959
UNION	EAGLE2.4.1	Kpbwt20000	18	0.5458
UNION	EAGLE2.4.1	Kpbwt20000	19	1.6468
UNION	EAGLE2.4.1	Kpbwt20000	20	0.7415
UNION	EAGLE2.4.1	Kpbwt20000	21	0.6411
UNION	EAGLE2.4.1	Kpbwt20000	22	1.0573
UNION	EAGLE2.4.1	Kpbwt10000	1	0.8792
UNION	EAGLE2.4.1	Kpbwt10000	2	0.6135
UNION	EAGLE2.4.1	Kpbwt10000	3	0.6839
UNION	EAGLE2.4.1	Kpbwt10000	4	0.7152
UNION	EAGLE2.4.1	Kpbwt10000	5	0.655
UNION	EAGLE2.4.1	Kpbwt10000	6	1.0743
UNION	EAGLE2.4.1	Kpbwt10000	7	0.6146
UNION	EAGLE2.4.1	Kpbwt10000	8	0.5107
UNION	EAGLE2.4.1	Kpbwt10000	9	0.76
UNION	EAGLE2.4.1	Kpbwt10000	10	0.6559
UNION	EAGLE2.4.1	Kpbwt10000	11	0.4801
UNION	EAGLE2.4.1	Kpbwt10000	12	0.655
UNION	EAGLE2.4.1	Kpbwt10000	13	0.5703
UNION	EAGLE2.4.1	Kpbwt10000	14	0.6778
UNION	EAGLE2.4.1	Kpbwt10000	15	0.9686
UNION	EAGLE2.4.1	Kpbwt10000	16	0.7544
UNION	EAGLE2.4.1	Kpbwt10000	17	1.1607
UNION	EAGLE2.4.1	Kpbwt10000	18	0.6989
UNION	EAGLE2.4.1	Kpbwt10000	19	2.0215
UNION	EAGLE2.4.1	Kpbwt10000	20	1.0381
UNION	EAGLE2.4.1	Kpbwt10000	21	0.7111
UNION	EAGLE2.4.1	Kpbwt10000	22	1.3792
UNION	CONSENSUS	High Resolution	1	0.5839
UNION	CONSENSUS	High Resolution	2	0.4359
UNION	CONSENSUS	High Resolution	3	0.4533
UNION	CONSENSUS	High Resolution	4	0.5098
UNION	CONSENSUS	High Resolution	5	0.5146
UNION	CONSENSUS	High Resolution	6	0.4986
UNION	CONSENSUS	High Resolution	7	0.4738
UNION	CONSENSUS	High Resolution	8	0.4345
UNION	CONSENSUS	High Resolution	9	0.6002
UNION	CONSENSUS	High Resolution	10	0.422

UNION	CONSENSUS	High Resolution	11	0.382
UNION	CONSENSUS	High Resolution	12	0.438
UNION	CONSENSUS	High Resolution	13	0.4393
UNION	CONSENSUS	High Resolution	14	0.4641
UNION	CONSENSUS	High Resolution	15	0.726
UNION	CONSENSUS	High Resolution	16	0.6417
UNION	CONSENSUS	High Resolution	17	0.8431
UNION	CONSENSUS	High Resolution	18	0.6088
UNION	CONSENSUS	High Resolution	19	1.6322
UNION	CONSENSUS	High Resolution	20	0.6715
UNION	CONSENSUS	High Resolution	21	0.6006
UNION	CONSENSUS	High Resolution	22	1.263
UNION	CONSENSUS	Default	1	0.6303
UNION	CONSENSUS	Default	2	0.501
UNION	CONSENSUS	Default	3	0.5275
UNION	CONSENSUS	Default	4	0.5381
UNION	CONSENSUS	Default	5	0.5773
UNION	CONSENSUS	Default	6	0.5647
UNION	CONSENSUS	Default	7	0.4994
UNION	CONSENSUS	Default	8	0.5009
UNION	CONSENSUS	Default	9	0.6272
UNION	CONSENSUS	Default	10	0.5415
UNION	CONSENSUS	Default	11	0.428
UNION	CONSENSUS	Default	12	0.4928
UNION	CONSENSUS	Default	13	0.4591
UNION	CONSENSUS	Default	14	0.555
UNION	CONSENSUS	Default	15	0.7749
UNION	CONSENSUS	Default	16	0.7119
UNION	CONSENSUS	Default	17	0.919
UNION	CONSENSUS	Default	18	0.6431
UNION	CONSENSUS	Default	19	1.7135
UNION	CONSENSUS	Default	20	0.8816
UNION	CONSENSUS	Default	21	0.6485
UNION	CONSENSUS	Default	22	1.3703
TWOSTAGE	BEAGLE5	PhaseStates280	1	0.2238
TWOSTAGE	BEAGLE5	PhaseStates280	2	0.1769
TWOSTAGE	BEAGLE5	PhaseStates280	3	0.184
TWOSTAGE	BEAGLE5	PhaseStates280	4	0.1956
TWOSTAGE	BEAGLE5	PhaseStates280	5	0.1721
TWOSTAGE	BEAGLE5	PhaseStates280	6	0.1344
TWOSTAGE	BEAGLE5	PhaseStates280	7	0.1912
TWOSTAGE	BEAGLE5	PhaseStates280	8	0.1689
TWOSTAGE	BEAGLE5	PhaseStates280	9	0.249
TWOSTAGE	BEAGLE5	PhaseStates280	10	0.2176
TWOSTAGE	BEAGLE5	PhaseStates280	11	0.1461
TWOSTAGE	BEAGLE5	PhaseStates280	12	0.1705
TWOSTAGE	BEAGLE5	PhaseStates280	13	0.2197
TWOSTAGE	BEAGLE5	PhaseStates280	14	0.2329

TWOSTAGE	BEAGLE5	PhaseStates280	15	0.3285
TWOSTAGE	BEAGLE5	PhaseStates280	16	0.2662
TWOSTAGE	BEAGLE5	PhaseStates280	17	0.3636
TWOSTAGE	BEAGLE5	PhaseStates280	18	0.3278
TWOSTAGE	BEAGLE5	PhaseStates280	19	0.973
TWOSTAGE	BEAGLE5	PhaseStates280	20	0.344
TWOSTAGE	BEAGLE5	PhaseStates280	21	0.2763
TWOSTAGE	BEAGLE5	PhaseStates280	22	0.7552
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	1	0.1629
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	2	0.1197
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	3	0.1216
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	4	0.1213
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	5	0.12
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	6	0.0756
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	7	0.1218
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	8	0.1074
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	9	0.2067
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	10	0.142
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	11	0.1185
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	12	0.1374
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	13	0.1366
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	14	0.1531
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	15	0.2207
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	16	0.2368
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	17	0.2937
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	18	0.2414
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	19	0.7987
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	20	0.3028
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	21	0.2432
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth8	22	0.6348
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	1	0.5945
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	2	0.4024
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	3	0.473
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	4	0.4956
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	5	0.4323
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	6	0.8187
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	7	0.4871
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	8	0.4062
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	9	0.5579
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	10	0.469
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	11	0.3697
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	12	0.4732
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	13	0.4239
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	14	0.5056
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	15	0.7546
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	16	0.5225
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	17	0.7592
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	18	0.5332

TWOSTAGE	EAGLE2.4.1	Kpbwt10000	19	1.3331
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	20	0.6612
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	21	0.5564
TWOSTAGE	EAGLE2.4.1	Kpbwt10000	22	1.0004
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	1	0.1953
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	2	0.123
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	3	0.1303
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	4	0.1213
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	5	0.1183
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	6	0.1006
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	7	0.1379
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	8	0.1211
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	9	0.2173
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	10	0.1492
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	11	0.1318
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	12	0.1663
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	13	0.1521
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	14	0.1691
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	15	0.2476
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	16	0.2368
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	17	0.3356
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	18	0.254
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	19	0.9062
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	20	0.3543
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	21	0.2653
TWOSTAGE	SHAPEIT4.1.2	PbwtDepth4	22	0.6304
TWOSTAGE	BEAGLE5	PhaseStates560	1	0.2258
TWOSTAGE	BEAGLE5	PhaseStates560	2	0.1631
TWOSTAGE	BEAGLE5	PhaseStates560	3	0.1824
TWOSTAGE	BEAGLE5	PhaseStates560	4	0.1965
TWOSTAGE	BEAGLE5	PhaseStates560	5	0.1818
TWOSTAGE	BEAGLE5	PhaseStates560	6	0.1226
TWOSTAGE	BEAGLE5	PhaseStates560	7	0.1827
TWOSTAGE	BEAGLE5	PhaseStates560	8	0.1728
TWOSTAGE	BEAGLE5	PhaseStates560	9	0.2478
TWOSTAGE	BEAGLE5	PhaseStates560	10	0.2125
TWOSTAGE	BEAGLE5	PhaseStates560	11	0.1675
TWOSTAGE	BEAGLE5	PhaseStates560	12	0.1849
TWOSTAGE	BEAGLE5	PhaseStates560	13	0.1986
TWOSTAGE	BEAGLE5	PhaseStates560	14	0.2281
TWOSTAGE	BEAGLE5	PhaseStates560	15	0.3133
TWOSTAGE	BEAGLE5	PhaseStates560	16	0.2727
TWOSTAGE	BEAGLE5	PhaseStates560	17	0.3396
TWOSTAGE	BEAGLE5	PhaseStates560	18	0.308
TWOSTAGE	BEAGLE5	PhaseStates560	19	0.8887
TWOSTAGE	BEAGLE5	PhaseStates560	20	0.3563
TWOSTAGE	BEAGLE5	PhaseStates560	21	0.2506
TWOSTAGE	BEAGLE5	PhaseStates560	22	0.7158

TWOSTAGE	CONSENSUS	Default	1	0.1768
TWOSTAGE	CONSENSUS	Default	2	0.1354
TWOSTAGE	CONSENSUS	Default	3	0.1366
TWOSTAGE	CONSENSUS	Default	4	0.1443
TWOSTAGE	CONSENSUS	Default	5	0.1192
TWOSTAGE	CONSENSUS	Default	6	0.1035
TWOSTAGE	CONSENSUS	Default	7	0.1427
TWOSTAGE	CONSENSUS	Default	8	0.1377
TWOSTAGE	CONSENSUS	Default	9	0.1973
TWOSTAGE	CONSENSUS	Default	10	0.1604
TWOSTAGE	CONSENSUS	Default	11	0.1389
TWOSTAGE	CONSENSUS	Default	12	0.1488
TWOSTAGE	CONSENSUS	Default	13	0.1676
TWOSTAGE	CONSENSUS	Default	14	0.1802
TWOSTAGE	CONSENSUS	Default	15	0.2544
TWOSTAGE	CONSENSUS	Default	16	0.2384
TWOSTAGE	CONSENSUS	Default	17	0.2857
TWOSTAGE	CONSENSUS	Default	18	0.2612
TWOSTAGE	CONSENSUS	Default	19	0.8742
TWOSTAGE	CONSENSUS	Default	20	0.2987
TWOSTAGE	CONSENSUS	Default	21	0.2469
TWOSTAGE	CONSENSUS	Default	22	0.6742
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	1	0.4952
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	2	0.3597
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	3	0.3972
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	4	0.4505
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	5	0.3787
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	6	0.6043
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	7	0.4224
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	8	0.3496
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	9	0.4757
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	10	0.4148
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	11	0.3003
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	12	0.3916
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	13	0.3929
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	14	0.4896
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	15	0.662
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	16	0.4572
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	17	0.5854
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	18	0.4359
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	19	1.2866
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	20	0.5891
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	21	0.5195
TWOSTAGE	EAGLE2.4.1	Kpbwt20000	22	0.8756
TWOSTAGE	CONSENSUS	High Resolution	1	0.1874
TWOSTAGE	CONSENSUS	High Resolution	2	0.1315
TWOSTAGE	CONSENSUS	High Resolution	3	0.1406
TWOSTAGE	CONSENSUS	High Resolution	4	0.1513

TWOSTAGE	CONSENSUS	High Resolution	5	0.1324
TWOSTAGE	CONSENSUS	High Resolution	6	0.0991
TWOSTAGE	CONSENSUS	High Resolution	7	0.1417
TWOSTAGE	CONSENSUS	High Resolution	8	0.1357
TWOSTAGE	CONSENSUS	High Resolution	9	0.2161
TWOSTAGE	CONSENSUS	High Resolution	10	0.1584
TWOSTAGE	CONSENSUS	High Resolution	11	0.1307
TWOSTAGE	CONSENSUS	High Resolution	12	0.1426
TWOSTAGE	CONSENSUS	High Resolution	13	0.1591
TWOSTAGE	CONSENSUS	High Resolution	14	0.1659
TWOSTAGE	CONSENSUS	High Resolution	15	0.2544
TWOSTAGE	CONSENSUS	High Resolution	16	0.2286
TWOSTAGE	CONSENSUS	High Resolution	17	0.2817
TWOSTAGE	CONSENSUS	High Resolution	18	0.2612
TWOSTAGE	CONSENSUS	High Resolution	19	0.8394
TWOSTAGE	CONSENSUS	High Resolution	20	0.2925
TWOSTAGE	CONSENSUS	High Resolution	21	0.2211
TWOSTAGE	CONSENSUS	High Resolution	22	0.6195

DECLARATION OF CO-AUTHORSHIP

The declaration is for PhD students and must be completed for each conjointly authored article. Please note that if a manuscript or published paper has ten or less co-authors, all co-authors must sign the declaration of co-authorship. If it has more than ten co-authors, declarations of co-authorship from the corresponding author(s), the senior author and the principal supervisor (if relevant) are a minimum requirement.

1. Declaration by	
Name of PhD student	Vivek Appadurai
E-mail	vivek.appadurai@regionh.dk
Name of principal supervisor	Dr. Thomas Werge
Title of the PhD thesis	Genetic Analysis of Complex Traits in Population Scale Datasets

2. The declaration applies to the following article	
Title of article	Legacy data, whole genome imputation and the analysis of complex traits: Lessons from the iPSYCH case-cohort study
Article status	
Published ☐ Date:	Accepted for publication ☐ Date:
Manuscript submitted ☐ Date:	Manuscript not submitted ■
If the article is published or accepted for publication, please state the name of journal, year, volume, page and DOI (if you have the information).	

3. The PhD student's contribution to the article ***(please use the scale A-F as benchmark)*** Benchmark scale of the PhD-student's contribution to the article **A**. Has essentially done all the work (> 90 %) **B**. Has done most of the work (60-90 %) **C**. Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
1. Formulation/identification of the scientific problem	C
2. Development of the key methods	B
3. Planning of the experiments and methodology design and development	C
4. Conducting the experimental work/clinical studies/data collection/obtaining access to data	C

3. The PhD student's contribution to the article ***(please use the scale A-F as benchmark)*** Benchmark scale of the PhD-student's contribution to the article **A**. Has essentially done all the work (> 90 %) **B**. Has done most of the work (60-90 %) **C**. Has contributed considerably (30-60 %) **D.** Has contributed (10-30 %) **E.** No or little contribution (<10 %) **F.** Not relevant	**A, B, C, D, E, F**
5. Conducting the analysis of data	A
6. Interpretation of the results	C
7. Writing of the first draft of the manuscript	B
8. Finalisation of the manuscript and submission	B

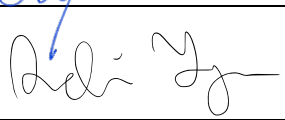
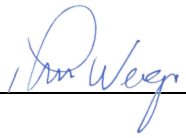
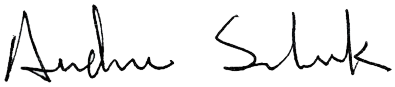
Provide a short description of the PhD student´s specific contribution to the article.[i]

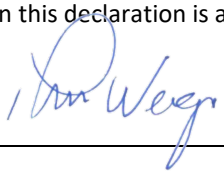
Contributed to the literature survey and planning of the study, performed the quality control, data analysis, wrote the manuscript and contributed to the final draft.

4. Material from another thesis / dissertation[ii]	
Does the article contain work which has also formed part of another thesis, e.g. master's thesis, PhD thesis or doctoral dissertation (the PhD student's or another person's)?	Yes: ☐ No: ■
If yes, please state name of the author and title of thesis / dissertation.	
If the article is part of another author's academic degree, please describe the PhD student´s and the author´s contributions to the article so that the individual contributions are clearly distinguishable from one another.	

5. Signatures of the co-authors[iii]

	Date	Name	Title	Signature
1.		Anders Rosengren	PhD	
2.		Jonas Grauholm	PhD	

5. Signatures of the co-authors[iii]				
3.		Alfonso Buil	PhD	
4.	14/12-2020	Andres Ingason	PhD	
5.		Thomas Werge	PhD	
6.		Olivier Delaneau	PhD	
7.		Andrew J Schork	PhD	
8.				
9.				
10.				

6. Signature of the principal supervisor
I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge. Date: 14/12-2020 Principal supervisor: 

7. Signature of the PhD student
I solemnly declare that the information provided in this declaration is accurate to the best of my knowledge. Date: 12-10-2020 PhD student: VIVEK APPADURAI 

Please learn more about responsible conduct of research on the Faculty of Health and Medical Sciences' website.

[i] This can be supplemented with an additional letter if needed.

[ii] Please see Ministerial Order on the PhD Programme at the Universities and Certain Higher Artistic Educational Institutions (PhD Order) § 12 (4):

"Any articles included in the thesis may be written in cooperation with others, provided that each of the co-authors submits a written declaration stating the PhD student's or the author's contribution to the work."

[iii] If more signatures are needed please add an extra sheet.