THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

The Genetic Architecture of Psychiatric Disorders

Robert Maier

Doctor of Medicine, Master of Science

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

The Queensland Brain Institute

Abstract

The genetic nature of psychiatric disorders was observed by clinicians long before DNA had been identified as the molecule of inheritance. The greatest identified risk factor of many psychiatric disorders still is a positive family history. Until recently this knowledge has not contributed substantially to treatment efforts or to a better understanding of the disease processes because we lacked the necessary genetic data. Advances in genotyping technologies have brought an end to this data shortage which is leading to a better understanding of the genetic architecture of psychiatric disorders. Two patterns started to emerge which were uncommon in earlier studied Mendelian disorders. (i) most of the genetic part of disease risk is conferred by a large number of genetic loci of small effect, and (ii) genetic loci often influence a large number of traits at the same time. While this is true of many traits ("complex" traits), these two phenomena (polygenicity and pleiotropy) are particularly pronounced in psychiatric disorders. This has wide-reaching consequences for the analysis and interpretation of genetic data and provides challenges as well as opportunities. This thesis focuses on two areas in particular: genetic heterogeneity and genetic risk prediction.

Genetic heterogeneity in a phenotypically homogenous group describes a situation where different and distinct genetic risk profiles are causing similar symptoms in different people. This can be easily identified under a Mendelian inheritance pattern, but proves to be challenging under polygenicity. The presence of genetic heterogeneity can limit the accuracy of genetic risk prediction.

The aim of genetic risk prediction is to use the information that has been gathered on the effects of genetic loci to estimate the genetic liability of an individual to develop a disease. Here, pleiotropy offers an opportunity to increase the accuracy of genetic prediction by leveraging information from multiple diseases at the same time.

The aim in this thesis is to describe several projects which center around the two concepts of genetic risk prediction based on multiple traits and of genetic heterogeneity. Chapter 1 sets the scene with an overview of recently developed polygenic methods. Chapter 2 deals with the effects of genetic heterogeneity on heritability estimates and demonstrates how genetic heterogeneity might contribute to the phenomenon of missing heritability, which is

the discrepancy between twin study heritability estimates and the variance explained by the sum of individual genetic loci. Chapter 3 addresses the question of whether genotype clustering can detect groups with different genetic risk profiles. Chapter 4 describes the implementation of a multivariate extension to the univariate Best Linear Unbiased Prediction (BLUP) method and its application to five psychiatric traits. Chapters 4 and 5 investigate whether this multivariate BLUP model can be approximated when only summary statistics, not individual level genotype data, are available for the predicted traits. Theory is derived for such an approximation, which is then tested in a simulation setup and applied to two psychiatric disorders, as well as to a range of other traits.

Finally, the discussion places the work into wider context and discusses the findings and limitations of each project, and highlights similarities and differences between the two prediction projects.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

First-author peer-reviewed papers

Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, William Coryell, James B. Potash, William A. Scheftner, Jianxin Shi, Myrna M. Weissman, Christina M. Hultman, Mikael Landén, Douglas F. Levinson, Kenneth S. Kendler, Jordan W. Smoller, Naomi R. Wray and Sang Hong Lee: **Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder**. The American Journal of Human Genetics (2015): 283-294

Non-first-author peer-reviewed papers

Naomi R Wray & **Robert Maier**: **Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability**. Current Epidemiology Reports 1.4 (2014): 220-227.

SM Meier, E Agerbo, **R Maier**, CB Pedersen, M Lang, J Grove, MV Hollegaard, D Demontis, BB Trabjerg and C Hjorthøj: **High loading of polygenic risk in cases with chronic schizophrenia.** Molecular Psychiatry 21 (2015): 969-974

Gwyneth Zai, Bonnie Alberry, Janine Arloth, Zsófia Bánlaki, Cristina Bares, Erik Boot, Caroline Camilo, Kartikay Chadha, Qi Chen, Christopher B Cole, Katherine T Cost, Megan Crow, Ibene Ekpor, Sascha B Fischer, Laura Flatau, Sarah Gagliano, Umut Kirli, Prachi Kukshal, Viviane Labrie, Maren Lang, Tristram A Lett, Elisabetta Maffioletti, **Robert Maier**, Marina Mihaljevic, Kirti Mittal, Eric T Monson, Niamh L O'Brien, Søren D Østergaard, Ellen Ovenden, Sejal Patel, Roseann E Peterson, Jennie G Pouget, Diego L Rovaris, Lauren Seaman, Bhagya Shankarappa, Fotis Tsetsos, Andrea Vereczkei, Chenyao Wang, Khethelo Xulu, Ryan K C Yuen, Jingjing Zhao, Clement C Zai & James L Kennedy: **Rapporteur summaries of plenary, symposia, and oral sessions from the XXIIIrd World Congress of Psychiatric Genetics Meeting in Toronto, Canada, 16-20 October 2015.** Psychiatric Genetics (2016): 229-257.

Conference abstracts

Using pleiotropy to improve genetic risk prediction in psychiatric disorders
Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, Cross disorder Working group of the Psychiatric Genomics Consortium, William A Coryell, James B Potash, William A Scheftner, Jianxin Shi, Myrna M Weissman, Christina M Hultman, Mikael Landén, Douglas F Levinson, Kenneth S Kendler, Jordan W Smoller, Naomi R Wray, S Hong Lee
Society of Mental Health Research Conference, 2014, Adelaide, Australia

Multivariate Genetic Risk Scores Increase Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder
Robert Maier, Naomi R Wray, Peter M Visscher & Matthew R Robinson
World Congress of Psychiatric Genetics Meeting, 2015, Toronto, Canada

Multivariate Genetic Risk Scores Increase Risk Prediction Accuracy for a Wide Range of Traits.
Robert Maier, Naomi R Wray, Peter M Visscher & Matthew R Robinson
Behavior Genetics Association Meeting, 2016, Brisbane, Australia

Multivariate Genetic Risk Scores Can Increase Risk Prediction Accuracy for a Wide Range of Traits.
Robert Maier, Naomi R Wray, Peter M Visscher & Matthew R Robinson
International Genetic Epidemiology Society Meeting, 2016, Toronto, Canada

Multivariate Genetic Risk Scores Increase Risk Prediction Accuracy for a Wide Range of Traits.
Robert Maier, Naomi R Wray, Peter M Visscher & Matthew R Robinson
American Society of Human Genetics Annual Meeting, 2016, Vancouver, Canada

Publications included in this thesis

Naomi R Wray, & **Robert Maier**: **Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability**. Current Epidemiology Reports 1.4 (2014): 220-227. – incorporated as Chapter 2: Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability.

Contributor	Statement of contribution
Robert Maier (Candidate)	Conception and Design (10%) Analysis and Interpretation (50%) Drafting and Production (40%)
Naomi Wray	Conception and Design (90%) Analysis and Interpretation (50%) Drafting and Production (60%)

Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, William Coryell, James B Potash, William A Scheftner, Jianxin Shi, Myrna M Weissman, Christina M Hultman, Mikael Landén, Douglas F Levinson, Kenneth S Kendler, Jordan W Smoller, Naomi R Wray & Sang Hong Lee: **Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder**. The American Journal of Human Genetics (2015): 283-294 – incorporated as Chapter 4: Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder.

Contributor	Statement of contribution
Robert Maier	Analysis and Interpretation (75%) Drafting and Production (40%)
Gerhard Moser	Conception and Design (10%)
Guo-Bo Chen	Conception and Design (10%)
Stephan Ripke	Provided data (9%)
William Coryell	Provided data (9%)
James B Potash	Provided data (9%)
William a. Scheftner	Provided data (9%)
Jianxin Shi	Provided data (9%)
Myrna M Weissman	Provided data (9%)
Christina M Hultman	Provided data (9%)
Mikael Landén	Provided data (9%)
Douglas F Levinson	Provided data (9%)
Kenneth S Kendler	Provided data (9%)
Jordan W Smoller	Provided data (9%)
Naomi R Wray	Conception and Design (10%) Drafting and Production (10 %)
Sang Hong Lee	Conception and Design (70%) Analysis and Interpretation (25%) Drafting and Production (50%)

Contributions by others to the thesis

Several other people have contributed significantly to this thesis. First and foremost, my supervisors Hong Lee, Matt Robinson, Naomi Wray and Peter Visscher, and the co-authors listed on the included publications and manuscripts.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

First, I would like to thank my supervisors Hong Lee, Matt Robinson, Naomi Wray and Peter Visscher for their continuing support, patience and for their quick, thoughtful and honest feedback.

Second, I want to thank all group members of CNSG/PCTG not only for providing a lot of academic support, but for making my time in Brisbane very enjoyable.

Third, I want to thank UQ and QBI for supporting me financially through scholarships and IGES, WCPG, SMHR and SISG for supporting me with travel grants.

Fourth, I want to thank my family and friends for putting more effort into staying in touch than I did.

Finally, I want to acknowledge some of the unsung heroes who made this work possible: The people who develop the free software that I used on a daily basis, especially R and all its great libraries, and the people who spend time and effort to share their expertise and knowledge on websites like “Stack Overflow” and others in the Stack Exchange network.

Keywords

quantitative genetics, statistical genetics, psychiatry, schizophrenia, bipolar disorder

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060412, Quantitative Genetics, 100%

Fields of Research (FoR) Classification

FoR code: 0604, Genetics, 100%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADHD: Attention deficit hyperactivity disorder
ARIC: Atherosclerosis Risk in Communities
ASD: Autism spectrum disorder
AUC: Area under the curve
BIP: Bipolar disorder
BLUP: Best linear unbiased prediction
BLUE: Best linear unbiased estimator
CAD: Coronary artery disease
COJO: Conditional and join analysis
CNS: Central nervous system
CNV: Copy number variation
DNA: Deoxyribonucleic acid
GBLUP: Genomic best linear unbiased prediction
GCTA: Genome-wide complex trait analysis
GERA: Genetic Epidemiology Research on Aging
GRM: Genetic relatedness matrix
GWAS: Genome wide association study
HLA: Human leucocyte antigen
IBD: Identity by descent / Inflammatory bowel disease
IBS: Identity by state
IQ: Intelligence quotient
LD: Linkage disequilibrium
LDpred: Linkage disequilibrium prediction
LDSC: LD score regression
LMM: Linear mixed model
LRT: Likelihood ratio test
MAF: Minor allele frequency
MDD: Major depressive disorder
MDS: Multi-dimensional scaling
MHC: Major histocompatibility complex
MME: Mixed model equations
MND: Motor neuron disease

MR: Mendelian randomization

MT: Multi trait

MTGBLUP: Multi trait genomic BLUP

MTGREML: Multi trait genetic REML

NCP: Non centrality parameter

OLS: Ordinary least squares

PC: Principal component

PCA: Principal component analysis

PEV: Prediction error variance

PGC: Psychiatric genomics consortium

PICS: Probabilistic Identification of Causal SNPs

QC: Quality control

REML: Restricted (or residual) maximum likelihood analysis

ROC: Receiver operator characteristic

SAI: Schizophrenia / Autism / Intellectual disability

SBLUP: Summary statistics BLUP

SD: Standard deviation

SE: Standard error

SCZ: Schizophrenia

SMR: Summary statistics Mendelian randomization

SMTpred: Summary statistics multi trait prediction

SNP: Single nucleotide polymorphism

ST: Single trait

STBLUP: Single trait BLUP

TWAS: Transcriptome wide association study

UK: United Kingdom

UKB: UK Biobank

Introduction

Our understanding of what we now call psychiatric disorders has changed profoundly in the last two centuries. The notion that the brain underlies all kinds of cognitive and emotional processes has led to the idea that a mental disorder may be a manifestation of malfunctioning brain processes, analogous to how somatic disease is often a consequence of the malfunctioning of other organs. This started the field of biological psychiatry and with it the difficult search for somatic correlates of psychiatric disorders that may aid in diagnostic classifications and treatment.

After the concept of the biological origin of psychiatric disorders became accepted, a separate but related question was to what degree heritable factors contribute to variation in susceptibility to psychiatric disorders. The true but not very informative statement that both genetic and environmental factors play a role could be refined by twin and family studies which provided estimates of how much genetic factors contribute to each trait or disease. These estimates pointed to a large genetic component of many psychiatric traits [1], implying that it would be worthwhile to study this component in greater detail [2,3]. This has become possible in the last fifteen years through the reduction in the cost of genotyping technologies. We can now address a multitude of questions about how genetic factors contribute to disease risk, often starting with the estimation of effects of individual genetic loci [4].

The first chapter of this thesis gives an overview over recently emerged methods that tackle some of these questions, including estimation of heritability and genetic correlation, inferring causality of SNPs and of phenotypes on other phenotypes, polygenic risk prediction and detection of disease heterogeneity.

The question of disease heterogeneity is of particular interest in psychiatry. Psychiatry uses carefully drafted disease classifications, which are in turn based on various symptom configurations, but it is an open question if the separations imposed by these classifications are mirrored in a similar structure on the genetic level [2,5]. If that is not the case, our current diagnostic criteria are not perfectly aligned with underlying biology, and what we classify as one disease might in fact be genetically distinct groups of diseases. This kind of disease heterogeneity is the subject of two chapters of this thesis.

Chapter 2 explores how disease heterogeneity affects estimates of heritability from pedigree studies and estimates of SNP heritability. The difference between heritability estimates from pedigree studies and estimates of SNP heritability has been labelled missing heritability and has been the subject of much research [6–8]. While much of missing heritability can be explained by imperfect tagging of causal markers by genotyped markers, we show that disease heterogeneity can also account for some of the difference between the two types of estimates, as they are differently affected by genetic heterogeneity.

For this and many other reasons, it seems desirable to define groups of affected individuals who exhibit a genetically more homogeneous risk profile. One approach to this is to cluster individuals based on a certain set of SNPs. Chapter 3 investigates in a simulation setup the degree to which such genotype based clustering methods are able to distinguish between disease subgroups. The results of this analysis are sobering in that even with very large sample sizes, a good separation between subgroups may not be possible for polygenic disorders.

However, genetic heterogeneity doesn't only pose problems. A flip side of heterogeneity is that what we classify as separate diseases might in fact have a shared aetiology. On a genetic level this may manifest itself as a positive genetic correlation between the two diseases. This is common across most traits and diseases, but especially common among psychiatric disorders, and provides a unique opportunity to improve genetic risk prediction [9,10]. In polygenic traits, genetic risk prediction crucially depends on the accurate estimation of the effects of associated (and non-associated) markers [4]. These estimates can be made more accurate by combining data on multiple, genetically correlated diseases. This concept is the subject of chapters 4 and 5.

Chapter 4 introduces a method to use genotype and phenotype data from multiple traits to derive genetic risk predictors which are more accurate than their counterparts for single traits. The method is a multivariate extension of the GBLUP method which is widely used in genomic prediction. We show that the method increases prediction accuracy in five psychiatric disorders, and quantify the amount of increase in sample size that would be necessary to achieve a similar increase using the single trait approach.

This method requires individual level genotype data on all traits that are to be combined.

Access to individual level data is often restricted and analysing these data comes with a large computational burden. Many methods overcome these problem by using summary statistics from genome wide association studies (GWAS). We therefore set out to develop a method that is equivalent to the multivariate GBLUP method and requires only summary statistics. After deriving theory for such a summary statistics based method, we conducted extensive simulations to test its performance. We then went on to apply this method to schizophrenia and bipolar disorder, as well as to a wide range of other traits, often finding large increases in prediction accuracy.

Finally, the discussion chapter concludes with a critical reflection on the work presented in the results chapters, highlighting in particular limitations of the approaches and differences between the two multi trait risk prediction projects.

1

Chapter 1: Embracing polygenicity: A review of methods and tools for psychiatric genetics research

Chapter has been submitted to Psychological Medicine

Chapter 1: Embracing polygenicity: A review of methods and tools for psychiatric genetics research

Robert M Maier[1,2], Peter M Visscher[1,2], Matthew R Robinson[2,3] & Naomi R Wray[1,2]

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.

[2]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia.

[3]Department of Computational Biology, University of Lausanne, 1011 Lausanne, Switzerland

Abstract

The availability of genome-wide genetic data on hundreds of thousands of people has led to an equally rapid growth of the range of methodologies available to analyse these data. While the motivation for undertaking genome-wide association (GWA) studies is identification of genetic markers associated with complex disease, once generated these data can be used for many other analyses. GWA data have demonstrated that complex traits (including psychiatric traits) exhibit a highly polygenic genetic architecture, often with shared genetic risk factors across traits. New methods to analyse GWA data are increasingly being used in research studies of psychiatric disorders, as well as many other fields of medicine, to address a diverse set of questions about the aetiology of complex traits and diseases. Here, we give an overview of some of these methods and present examples of how they have contributed to our understanding of psychiatric disorders. The methods are concerned with (i) estimation of the extent of genetic influence on traits, (ii) uncovering of shared genetic control between traits, (iii) predictions of genetic risk for individuals, (iv) uncovering of causal relationships between traits, (v) identifying causal SNPs and genes or (vi) the detection of genetic heterogeneity. This classification helps to organise the large number of recently developed methods, however some of them could be placed in more than one of these classes. While some methods require GWA data on individual people, others simply use GWA summary statistics data, allowing novel well-powered analyses to be conducted at a low computational burden.

Introduction

The reduction in costs of genotyping technologies in recent years has led to an explosion of genetic and phenotypic information collected on large numbers of people. The primary aim of these studies is to find genetic polymorphisms associated with a quantitative trait or with an increased risk of disease. These genome-wide association studies (GWAS) have led to an important increase in understanding of the underpinnings of psychiatric and other disorders [3,11,12]. However, the potential use of these data goes far beyond merely mapping genetic variation to disease.

Here, we present an overview of some recently developed methods utilizing genome-wide-genotype and phenotype data on large numbers of individuals and show how they can be applied to the research of psychiatric disorders. These new methods serve at least one of the following purposes: (i) estimation of the extent of genetic influence on traits (**Estimation of proportion of variance attributable to genome-wide single nucleotide polymorphisms (SNPs)**, SNP-heritability or h^2_{SNP}), (ii) uncovering of shared genetic control between traits (**Estimation of genetic correlation from using genome-wide SNPs**), (iii) predictions of genetic risk for individuals (**Polygenic risk prediction**), (iv) uncovering of causal relationships between traits (**Mendelian randomization**), (v) identifying causal SNPs and genes (**Fine-mapping and gene prioritization**), or (vi) **Detection of genetic heterogeneity**. There is often an overlap between these applications: some methods can be applied for more than one purpose, and some of the available software implements more than one method. Other applications of GWAS summary statistics such as pathway analysis [13] are outside the scope of this review.

Most of the methods presented here require genetic data in one of two possible formats. The first data format is that of full individual level genotype data and phenotypic measurements on each person, where the genetic data can be represented as a matrix with allele counts for each genetic marker for each person. While this offers the largest range of analytic options, file sizes can be very large, which can become prohibitive as computational burden is usually non-linear with increasing numbers of individuals and markers. Moreover, privacy concerns can prevent this type of data from being shared across research groups. Summary statistics of genome-wide association analysis represent the second data format, for which data sharing has fewer privacy concerns [4]. GWAS summary statistics comprise the association test statistic (including direction of effect for a reference allele), standard error, p-value of association and allele frequency of each SNP. While it has been shown that it is possible to infer whether an individual was part of a cohort using summary statistics, the power to do so is limited [14–16], and in any case requires the genome-wide genotype data of the individual to be identified. To guard against any privacy concerns GWAS summary statistics can be provided using allele frequencies estimated in large independent samples of the same ethnicity. Methods which require only summary statistic data benefit from shorter analytical run-times, much reduced computer memory requirements, and applicability to a larger number of traits.

While genetic data in one of these two formats are required by all methods presented here, some methods additionally make use of other information such as genomic annotation and expression quantitative trait locus (eQTL) data. Genomic annotation can give clues about the functional importance of a region in which a SNP resides, whereas eQTL data are the result of an association test where the phenotype of interest is the expression of a particular gene.

Here, we review a range of different polygenic methods and highlight their aims and the input data they require (**Figure 1**). For each method we provide some examples of applications relevant to psychiatric genetics research.

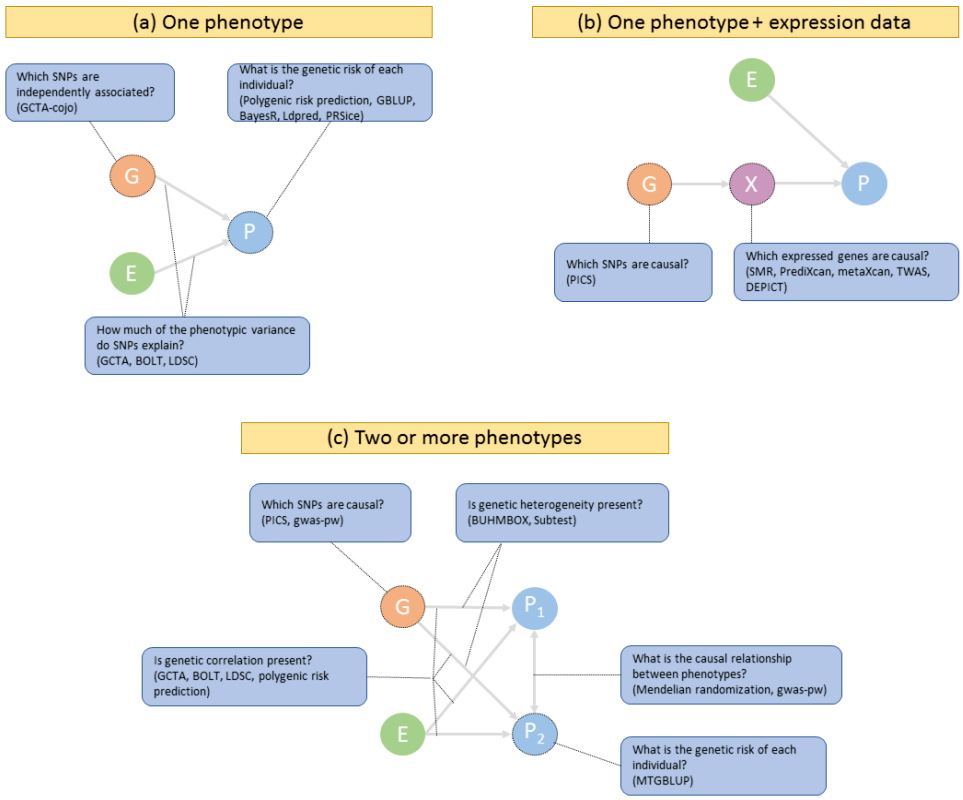

Figure 1: Schematic of the basic models underlying the polygenic methods reviewed
All models assume that a phenotype (P) is influenced by genetic (G) and environmental (E) factors (with environmental defined loosely as anything not captured by G including stochastic variation and measurement error). (a) model which considers only one phenotype

and no gene expression. (b) model which considers only one phenotype and gene expression (X). (c) model which considers two or more phenotypes and no gene expression. Methods can be grouped into those where the focus lies on individual SNPs, genes or people (nodes highlighted), and those where the focus lies on aggregate measures affecting the relationship between genetic and environmental factors and a phenotype (edges highlighted).

Estimation of proportion of variance attributable to genome-wide SNPs

Heritability is the proportion of phenotypic variance that can be attributed to genetic factors. It is a key quantity in genetics research as it summarizes the role of causal inherited variation. Trait heritability can be estimated by comparing the phenotypic resemblance among family members to their coefficients of relationships. However, estimating heritability from relatives can result in upwards-biased estimates [17] if non-genetic factors shared by relatives cannot be disentangled from the shared genetic relationships. This can be circumvented by estimating heritability from genome-wide markers in unrelated individuals. Genomic restricted maximum likelihood analysis (GREML) can be used to estimate the proportion of phenotypic variance which is captured by genotyped SNPs (SNP-heritability), by using genetic data of unrelated individuals [18]. SNP-heritability estimates are typically lower than from twin or family based heritability estimates, because genotyped SNPs account only for a subset of all genetic effects (the remainder includes other types of polymorphisms and SNPs that are not tagged by genotyped SNPs). Hence, the parameter estimated by SNP-heritability analysis depends on the genotype data available in the data set analysed, and can only converge to the traditional parameter being estimated from family data, when the genotypes available are fully representative of the variation in the genome.

Estimation of SNP-heritability has been of particular importance for disease traits, especially those of low lifetime risk (<1% is typical of most common diseases) for which it is difficult to collect the large samples needed to calculate heritability from estimates of increased risk in relatives of those affected. Both traditional-heritability and SNP-heritability estimates are presented on the liability scale (and depend on lifetime risk of disease in the population), and empirical data of the GWAS era [19,20] demonstrates that the polygenic model implied in these estimates is justified. GWAS case-control samples for disease traits are usually heavily over-sampled for cases compared to a population sample and so SNP-heritability estimates are made on this binary case-control scale and transformed to the liability scale accounting for this ascertainment [21]. The SNP-heritability estimates are relatively robust to choice of lifetime risk for most common diseases (lifetime risk < 1%). However, for the very common diseases such as major depressive disorder (lifetime risk 15%), SNP-heritability estimates are more sensitive to the choice of lifetime risk estimate and to screening vs non-screening of controls ([22], **Figure 2**). Furthermore, the ascertainment of cases in case-control studies can induce an artificial gene-environment correlation, which

may lead to underestimation of heritability, especially in rare diseases, where the ascertainment of cases is most extreme. It has therefore been recommended to instead use another method, Haseman-Elston (H-E) regression to estimate heritability in case-control studies [23]. In H-E regression, the heritability estimate is obtained by regressing pairwise phenotypic similarity on pairwise genetic similarity. This corresponds to a model where the phenotypic similarity is in expectation equal to the genetic similarity multiplied by the heritability of the trait. This model is most appropriate when the phenotype is an additive, polygenic, quantitative trait and the individuals comprise a random sample of the population. PCGC is a method which generalises H-E regression, by allowing more general models, in which the phenotypic similarity depends on heritability and genetic similarity in more flexible ways [23].

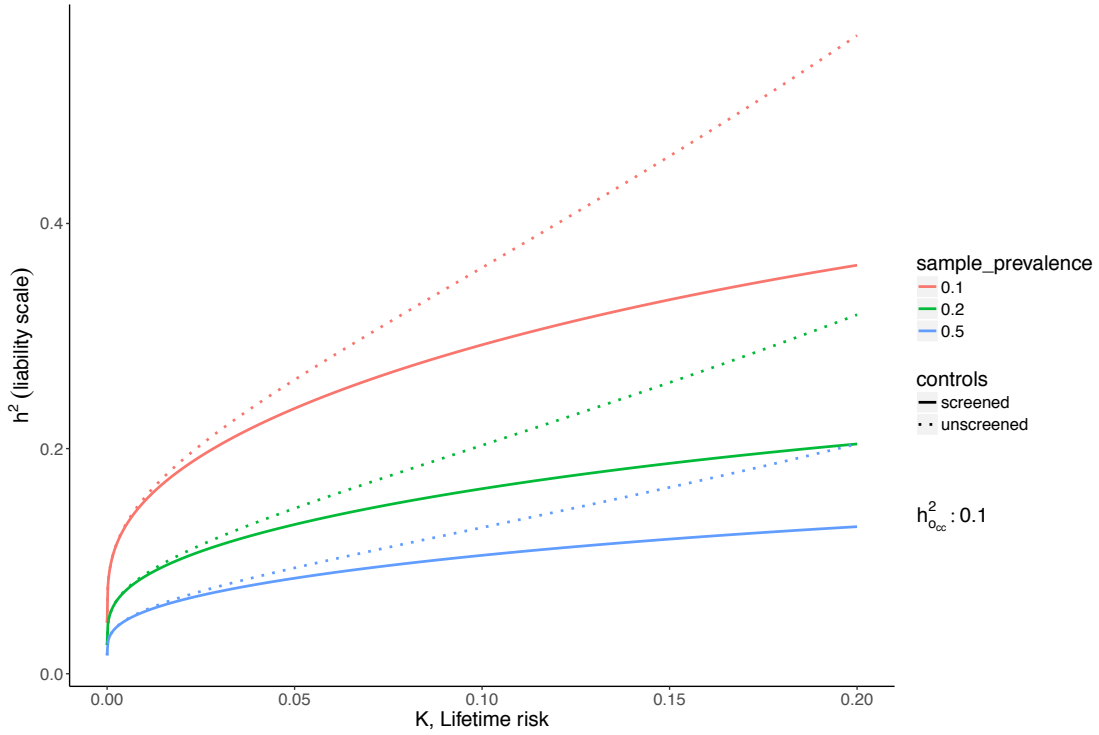

Figure 2: Conversion of heritability from the observed case/control scale to the liability scale

An estimate of SNP-heritability from a case-control sample reflects the properties of the sample as well as the properties of the disease. Here we show that an estimate of SNP-heritability of 0.1 from the linear model applied to a case-control sample (h^2_{occ}) can reflect a

very different heritability on the liability scale, depending on the proportion of cases in the sample (colours), the lifetime prevalence K (x-axis), and depending on whether or not controls were screened (solid lines (Lee et al. 2011, Eq. 23) vs dotted lines (Peyrot et al. 2016, Eq. 3)).

The GREML method for estimation of SNP-heritability is based on a linear mixed model [24], where a central component is the genetic relationship (or similarity) matrix (GRM), which captures the genetic relatedness of all pairs of individuals (Hayes et al., 2009). For application to human data, the algorithm is implemented in GCTA (Yang et al., 2011). Several independent studies have confirmed that GREML results in unbiased estimates of SNP heritability when the model assumptions are met: The model states that the phenotypic similarity between individuals can be decomposed into a genetic component, which is given by the genetic relationship matrix multiplied by the heritability of the phenotype, and a residual component of uncorrelated errors. It assumes that each SNP is causal and that the variance explained by a SNP is independent of its MAF. This is equivalent to rare SNPs having a larger effect size. Further assumptions are that SNP effects as well as random errors are normally distributed. It has been investigated how departures from these assumptions can influence the results [27,28]. Potential biases that might influence heritability estimation are an association between MAF and SNP effect size which doesn't fit the assumptions of the model, or an overrepresentation of causal SNPs in regions of high or low linkage disequilibrium (LD) [21,27]. To overcome the problem of bias introduced by an enrichment of causal variants in regions of high or low LD, it has been suggested to use an LD weighted GRM instead of the normal GRM, as implemented in the LDAK program [27]. Another solution to this problem is to stratify the GREML analysis by MAF and LD, as implemented in GCTA GREML-LDMS [8]. A recent comparison of multiple heritability estimation has shown that GREML-LDMS can overcome these biases and performs as well or better than other methods under most simulation scenarios [29]. The runtime of GCTA-GREML is a function of both the number of markers, M, and the number of individuals, N. Compute time is $O(N^3 + MN^2)$, which includes a component for the construction of the genetic relationship matrix and a component for the actual REML algorithm. The steep increase in runtime with larger N makes it impractical for data sets with very large numbers of individuals. The software BOLT-LMM can estimate variance components through a stochastic approximation algorithm, which circumvents the costly calculation of a genetic

relationship matrix and thus reduces the runtime to only $O(MN^{1.5})$ [30]. Note that both GCTA and BOLT-LMM have other features such as linear mixed model association analysis [31], which are not the topic of this review.

LD score regression (LDSC) [32] is a method which requires only summary statistics to estimate SNP-heritability and has therefore even shorter runtime than the methods discussed so far. Under polygenic genetic architecture, SNPs which are highly correlated with many other SNPs (have a high LD score) are more likely to tag a causal SNP and are therefore expected, on average, to have a higher association test statistic than SNPs which are not highly correlated with many other SNPs. The regression coefficient of the association test statistics of all SNPs on their LD score is a function of SNP-heritability [33,34]. Any factor that increases the association statistic of a SNP independently of its LD score (as might be found in population stratification which induces correlations in test statistics across chromosomes) will increase the intercept term of this regression [34]. LDSC estimates SNP-heritability with vastly reduced computational speed compared to GREML.

The standard error (s.e.) of GREML SNP-heritability estimates is accurately approximated as 316/N, where N is the total sample size [35]. For traits with very high or very low heritability the estimation of the standard error can be inaccurate. A more accurate bootstrap based method has been developed which yields unbiased standard errors, and thus confidence intervals, for GREML heritability estimates [36]. For LDSC the standard errors of the variance component estimates are typically larger (usually by 50% or more) than those of a GREML analysis for the same sample size [37]. However, it is typical that LDSC can be applied to larger data sets (which generate smaller s.e.) since only summary statistics are needed. Comparisons of estimates from GREML and LDSC show that the accuracy of estimates from LDSC are dependent on LD scores calculated from a population representative of the population used to estimate GWAS summary statistics [8,38].

Examples of applications to psychiatric disorders

GREML SNP-heritability estimated for psychiatric disorders usually ranges from 15% to 30%, depending on the disease, and these estimates are roughly half of the estimates derived from family studies [10,19,39]. *However, some studies estimate the SNP-heritability*

of Autism Spectrum Disorder to be between 50% - 60% [40,41]*, while other studies give estimates ranging from 17% - 24%* [10,42]*. These different estimates might result from differences in the ascertainment of cases and of controls between the studies. SNP-heritability estimates of quantitative mental traits tend to be relatively low (< 0.15), for example, a meta-analysis of the Big Five personality traits reported significant SNP-heritability estimates below 0.2 for all five traits* [43]*. This is in line with another study of up to 300,000 people finding SNP-heritability estimates for subjective well-being, depressive symptoms and neuroticism in the same range* [44]*. LDSC can also be used to estimate the SNP-heritability of specific genomic regions, such as enhancer regions. Furthermore, cell type-specific genomic annotations can be used to identify cell types or tissues which play a significant role in disease aetiology: In disease-relevant cell types, genomic annotations will more often overlap with causal SNPs than in other cell types. This type of analysis has identified the central nervous system as the most relevant tissue in the aetiology of schizophrenia and bipolar disorder* [45]*.*

Estimation of genetic correlation using genome-wide SNPs

Two traits are genetically correlated, if there is a correlation between the true effect sizes of SNPs affecting the two traits, or in other words, when, on-average, SNPs have directionally similar effects on two traits. For example, a genetic correlation of say zero, could imply no pleiotropy at all across the genome, or it could imply mixed directionality of pleiotropy. A positive genetic correlation estimate doesn't have to imply a correlation between the true effects for two traits, however. It can also occur when the causal SNPs for both traits are in LD with each other. This could for example happen after an admixture event where one population differs in both traits from the other population. Genetic correlations (r_G) are of interest because generally, they suggest a shared aetiology. However, they also be caused by misdiagnosis between two diseases [46,47]. The availability of genetic marker data on disease case-control samples has allowed the interrogation of the genetic relationship between diseases often for the first time, since traditional methods to estimate genetic correlation based on increased risk of a disease in relatives of those with another disease requires often unattainably large samples [35].

Formally, genetic correlation is defined as genetic covariance between two traits, scaled by the product of the genetic standard deviations of the two contributing traits. Methods used to estimate SNP heritability can be extended into a bivariate form to estimate r_G. In order to estimate genetic correlation using GREML, individual-level genetic data and measurements on two phenotypes are required (from the same or from different individuals). The power of bivariate GREML analyses to detect r_G departing from 0 or from 1 depends on the population value of r_G, the SNP-heritability of both traits, on the sample sizes, on whether the same or different samples are used for the two traits, and for disease traits, on the proportion of cases in the sample [35]. For example, for two diseases with lifetime prevalence of 1%, SNP-h^2 of 0.2 and genetic correlation of 0.5, 5000 cases and 5000 controls for each disease are sufficient to have 89% power at type 1 error rate of 5% to detect a genetic correlation greater than 0, corresponding to a standard error of 0.06. The BOLT-LMM software is also capable of calculating r_G in a bivariate GREML analysis with shorter runtime [30]. In contrast to heritability estimates, r_G estimates are scale independent (approximately) and hence scale transformation is not needed [48].

If summary statistics on two or more traits are available, LDSC can be used to estimate r_G between them, albeit with higher standard errors than GREML. Just as in the bivariate GREML analysis, sample overlap between the two traits should not affect the estimate of r_G. While it is easy to detect sample overlap in the presence of individual level data through the calculation of the GRM, it is much harder to detect if only SNP level summary statistics are available. When performing a bivariate LDSC analysis, the intercept of the regression is informative on whether the two sets of summary statistics are based on overlapping individuals [9].

To make the application of LDSC for SNP-heritability and r_G estimation even more user-friendly, the LD Hub resource has been developed, which provides access to summary statistics from more than 200 different traits and can calculate genetic correlation estimates of each of them with data provided by the user [49].

Estimation of r_G is most commonly applied to uncover genetic relationships between two different traits, but it can also be applied to detect heterogeneity in genetic effects between two different groups, for example a trait might be under different genetic control in men and in women or in old people and young people. It has also been applied to two data sets of

the same disease, where r_G should be one [10] to infer between-sample heterogeneity or two data sets of the same disease but of different ethnicity [50]. Calculating r_G across populations is not straight-forward, however, due to differences in both allele frequencies and LD structure. The program Popcorn addresses this problem and allows to estimate the transethnic genetic correlation based on summary statistics and LD matrices from two populations [38].

Examples of applications to psychiatric disorders

One of the first application of the bivariate GREML method to disease traits was to estimate genetic correlations between psychiatric disorders, presenting evidence that most pairs of disorders result in estimates that are significantly different from zero (PGC Cross disorder group, 2013)*. From those initial estimates the high correlations between schizophrenia and bipolar disorder (~0.6) and between bipolar and major depressive disorder (MDD) (0.5) were considered plausible given evidence from family studies, however the high genetic correlation between schizophrenia and MDD was more surprising for the clinical community. In fact, close study of the literature from family studies* (PGC Cross disorder group, 2013) *showed that all three genetic correlation estimates were consistent with published increased risk to relatives (RR) of one disorder for individuals diagnosed with another. However, for the same genetic correlation the size of RR involving a very common disorder (~15% lifetime risk) such as MDD is much smaller than when both disorders are less common (<1% lifetime risk for both schizophrenia and bipolar disorder). This connection between RR and genetic correlation is the bivariate analogue of the univariate case of computing heritability from risk to relatives. For example, when estimating heritability from first degree relatives, a disease with a lifetime risk of 15% and a relative risk to relatives of 1.5 has a heritability on the liability scale of 37%. For a disease with lifetime risk of only 1%, a relative risk to relatives of 3.17 would be needed to achieve the same heritability on the liability scale* [51]*.*

Genetic correlation can arise through misdiagnosis between two diseases. For example, those first presenting with clinical features consistent with a diagnosis of bipolar disorder can in the long-term receive a diagnosis of schizophrenia (and vice versa) [52,53]*. However, it can be shown analytically that very high misclassification rate of 20% would be needed under no shared aetiology to result in r_G of ~0.6 estimated between schizophrenia and bipolar disorder* [47]*. In contrast, a high genetic correlation between disorders would be consistent with some clinical presentations being difficult to classify.*

LDSC has been applied to a full battery of GWAS summary statistics and report higher estimates of genetic correlations estimates between pairs of psychiatric disorders than between psychiatric- and non-psychiatric disorders [9]*. Some notable examples include a positive genetic correlation between schizophrenia and anorexia nervosa, and a positive genetic correlation between bipolar disorder and years of education. A study investigating genetic sharing between neurological and psychiatric traits found that among neurological disorders significant genetic correlations are rare, but that there is some overlap in genetic risk between migraine and major depressive disorder, ADHD and Tourette syndrome* [39]*. On the other hand, genetic correlations between psychiatric traits and personality traits are more common.* **Figure 3** *shows the top genetic correlations for psychiatric disorders and traits. Data obtained from LD Hub.*

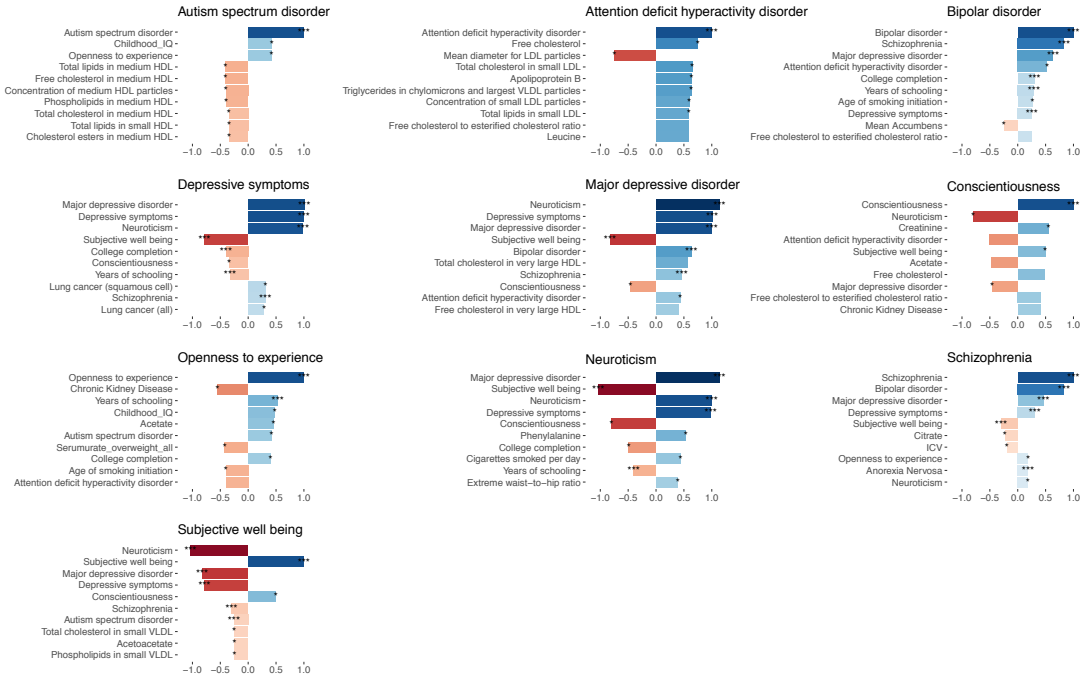

Figure 3: Genetic correlations between psychiatric disorders and traits, and almost 200 other traits

For each trait, the 10 traits with the highest absolute genetic correlations are shown. Colours indicate whether genetic correlations are positive or negative. One star indicates a genetic correlation p-value < 0.05. Three starts indicate a p-value below the Bonferroni threshold of 2.81×10^{-6} *for 17766 tested trait pairs. Data obtained from LD Hub* [49]*.*

Polygenic risk prediction

Estimates of SNP effects can be used to predict the genetic risk of individuals. Simple risk scores for each individual are calculated as the sum over all per SNP effects, where the per SNP effect is the allele count of the SNP for the individual multiplied by the effect size of the SNP [54,55]. Here the SNP effects come from a typically large and well powered discovery (sometimes called training) data set. While the ultimate goal of genetic risk prediction is in applications where the phenotype has not yet been observed, in research applications the risk predictor is evaluated for individuals in a target (sometimes called validation or testing) data set where the phenotype has already been recorded, so that the efficacy of the predictor can be evaluated.

Screening of high risk individuals for early intervention or prevention programs is a potential clinical application of polygenic risk prediction, that at present is not widely used because of the low accuracy of genetic risk predictors [56]. However, there are applications of polygenic risk prediction in research where low prediction accuracy is less limiting; for example, it could be a cost-effective strategy to conduct follow-up studies in samples ascertained to be low or high for polygenic risk. The genetic predictor is evaluated against a measured phenotype in the target sample, which may or may not be the same phenotype from which the predictor was constructed. Generally, if the predicted and the measured trait are genetically correlated, there should be a positive prediction accuracy, given enough power in both data sets [57].

In the usual implementation of polygenic risk scores, SNP effect sizes have been estimated from the standard, one SNP GWAS analyses. Then construction of polygenic risk scores is based on some decision about the proportion of SNPs to include in the predictor. As the discovery sample p-value threshold becomes more lenient the increased predictive power of including estimated effect sizes from more true positive associated SNPs is balanced by the inclusion of more false positives. The optimum proportion of SNPs to include depends on the (unknown) genetic architecture and size of the discovery sample. In the latest schizophrenia GWAS, the optimum p-value threshold was identified as 0.05 (based on variance explained in out-of-sample prediction across many samples), although inclusion of all SNPs did not lower accuracy drastically [58]. Often a range of different p-value thresholds are used to determine the best predictor, although this approach is prone to overfitting.

Software such as PLINK [59,60] implements basic polygenic risk scoring, while PRSice compares polygenic risk predictors using a large number of p-value cut-offs to find the optimum threshold given the data [61]. In polygenic risk scoring, LD among SNPs is usually accounted for by applying LD-clumping, i.e. pruning of SNPs based on LD but with higher preference to SNPs with lower p-values (usually simply termed “clumping”).

Prediction accuracy can be improved by using methods which provide estimates of SNP effects that are conditional on all other SNPs, thereby directly taking SNP LD correlations into account [62]. One example of such a method is Genomic Best Linear Unbiased Prediction (GBLUP) [24], which is widely used in animal breeding. In GWAS there are many more SNPs compared to individuals so effects sizes of each SNP are not estimable in a multiple regression model. In GBLUP it is assumed that SNP effect sizes are drawn from a normal distribution and a shrinkage term, or penalty term, proportional to the trait heritability is introduced in the model. This shrinks the SNP effect estimates (i.e., the effect attributable to SNPs in LD with each other is shared between the correlated SNPs) and will ensure the predicted phenotypes are on the right scale (unbiased) as well as being more accurate. More complex models, such as BayesR, remove the assumption of a single normal distribution of effect sizes, and allow for more general genetic architectures better by simultaneously estimating the true distribution of all SNP effect sizes and choosing individual SNP effect estimates accordingly [63]. Application of BayesR showed improved prediction over GBLUP for traits with some SNPs of larger effect (such as auto-immune disorders), but no improvement for other disorders (including bipolar disorder) [63].

There are a number of other prediction models which fit all SNPs at the same time and thereby utilize LD-information [64]. For example, LASSO (least absolute shrinkage and selection operator), which like GBLUP, shrinks SNP effects. In contrast to GBLUP, LASSO will shrink the effect of some SNPs to zero, and thus effectively select a subset of SNPs to include in the prediction. Underlying genetic architecture will dictate which method is better in practice. A study comparing the LASSO method to other prediction models found that it results in similar or higher prediction accuracy for immune disorders (an architecture that includes some SNPs of large effect), but not for bipolar disorder which is highly polygenic [65]. Finally, GBLUP approaches can be extended to a multi-trait model, which can further improve prediction accuracy when phenotypes are genetically correlated, because measurements on each trait provide information on the genetic values of the other correlated

traits. This approach has been implemented in the program MTG2 and has been shown to improve prediction accuracy for schizophrenia and bipolar disorder [66].

If individual-level genotype data are not available, it is still possible to transform marginal SNP effects (standard GWAS summary statistics) into penalized, conditional SNP effects, by making use of an LD reference data set [67,68]. This is implemented in GCTA [26] and in LDpred [69], which not only accounts for LD between SNPs, but also uses a Bayesian framework to adjust the SNP effects for traits where an infinitesimal model (all SNPs have some effect) is not the best fit to the data (e.g., autoimmune disorders). This is analogous to the selecting SNPs based on p-value in standard polygenic risk prediction and can further improve accuracy.

Examples of applications to psychiatric disorders

Polygenic risk prediction is widely used in psychiatric genetics, not to infer an individual's case control status, but to gain a better understanding of disease aetiology. Polygenic risk prediction in applications relevant to psychiatry has been reviewed previously [70]*. Some more recent examples include schizophrenia polygenic risk scores calculated for community samples of individuals which explain variation in creativity* [71] *and cannabis use* [72]*. An association between schizophrenia polygenic risk scores and negative symptoms and anxiety disorder in adolescents gives reason to hope that these kind of studies can not only lead to a better understanding of disease aetiology but may someday also contribute to early intervention programs* [73,74]*. Polygenic risk prediction additionally provides a novel approach to studying gene – environment interactions (GxE). Traditional GxE studies, which test for an interaction effect between single genetic variants and an environmental exposure on disease risk, often suffered from low power caused by the small amount of variance explained by individual genetic loci. In contrast, interactions between a polygenic risk score and environmental exposure can be detected more easily, because polygenic risk scores explain more of the variance in disease risk than individual loci. This type of GxE study has been applied to investigate a potential interaction between a polygenic risk score for major depressive disorder and childhood trauma on the risk for major depressive disorder. Two independent studies found that both the polygenic risk score and exposure to childhood trauma increase the risk for major depressive disorder, but came to different conclusions about the nature of the interaction between the two: One study found a positive interaction, meaning that those with both exposure to childhood trauma and high polygenic risk scores*

are at the greatest risk of developing major depressive disorder [75]*, while another study found a negative interaction, where people with exposure to childhood trauma and low polygenic risk scores are at the highest risk* [76]*. The latter study attributes these different findings to design differences between the two studies: The first study, which found a positive interaction effect, was a larger, population based study with a less stringent definition of depression and used a different instrument to assess childhood trauma.*

Mendelian randomization

Mendelian randomization (MR) analysis investigates the causal relationships between traits. It is a specific form of an instrumental variable analysis [77], where the goal is to test the causal effect of an explanatory variable (exposure to a risk factor) on a dependent outcome variable (such as disease risk). In MR, genetic markers are used as the instrumental variables. For example, MR was used to investigate a causal influence of HDL and LDL cholesterol levels on the risk for myocardial infarction [78]. They first identified SNPs which significantly lowered HDL cholesterol. If HDL cholesterol levels were causally related to the risk of myocardial infarction, these same SNPs should also be associated with a lower risk of myocardial infarction. However, no such association was found and so it was concluded that these data are inconsistent with a protective role of HDL for myocardial infarction. On the other hand, SNPs which increase LDL cholesterol were found to also increase the risk of myocardial infarction, confirming that LDL is a risk factor. Lowering LDL levels is a well-established intervention to reduce the risk of coronary artery disease [79], and the question of whether raising HDL levels can be similarly effective is of large public health interest. Together with a number of randomized controlled trials [80], this application of MR has had major impact on drug development by providing evidence against a causal role of HDL and thus helped in the search for effective ways of preventing myocardial infarction, and demonstrates how evidence from an MR analysis could be used to circumvent costly randomized controlled trials.

Despite its great potential, MR is often limited by low power, and by the fact that it is very difficult to show that all the assumptions which are necessary to infer causality are met. One of them is the absence of pleiotropy, since the SNPs used in the analysis may not have independent effects on the exposure and on the outcome. However, if MR is applied

bidirectionally for trait pairs of approximately comparable power, and evidence for significant causality is detected in only one direction, then this can help to infer causality over pleiotropy. The power in MR studies is a function of the true causal association between exposure and outcome and of the variance explained by the instrumental variables [81]. Since statistically significantly associated SNPs often only explain a small proportion of the genetic variance, for many pairs of traits, very large sample sizes are needed to achieve sufficient power to detect causal associations (see online calculator [81]).

In recent years, many improvements to the MR method have been developed. Apart from the extension to more than one instrumental variable (SNP), they include utilization of summary statistic data [82], better ways to test some the assumptions, modifications which allow the relaxation of the no-pleiotropy assumption, and improvements which increase the power to detect causal effects [77]. The web based resource MR BASE has been developed to simplify the application of MR to test causality between a large number of traits and to compare different variations of the method [83]. Summary data for more than a thousand traits have been collected and can be tested for causal associations with data provided by the user.

Examples of applications to psychiatric disorders

Several Mendelian randomization studies have investigated a potential causal influence of variables which are known to be associated with psychiatric traits and diseases from observational studies. One study which looked at BMI as a potential risk factor concluded that there is no evidence of BMI being a causal influence on schizophrenia and bipolar evidence, but weak evidence of BMI conferring a higher risk of major depressive disorder. It was noted, however, that the BMI association suffers from low power caused by small sample sizes for major depressive disorder [84]*. C-Reactive protein (CRP) is a potential risk factor for psychiatric disorders with undisputed correlational association, but unclear causality. A study from 2016 surprisingly found a protective role of genetically elevated CRP levels on the risk for schizophrenia, as well as weak (nominally significant) evidence for a risk increasing effect on bipolar disorder as well as a range of other somatic traits* [85]*. This is in contradiction with a another 2016 study which identified elevated CRP levels to confer an increased risk of schizophrenia* [86]*. There is much debate on whether cannabis is a risk factor for psychosis or schizophrenia, or whether the association is due to reverse causation or due to a confounding factor* [87,88]*. Recently MR studies have provided evidence for a*

causal role of cannabis in the development of schizophrenia, but also for a reverse causation [89,90]. *Several other risk factors for schizophrenia, anxiety and depression have been investigated through MR with negative results* [91–93].

In summary, MR studies investigating risk factors for psychiatric disorders could in a few cases provide evidence for a risk increasing effect. As the power of GWAS with increasing sample sizes, there will be more robust SNP associations which can be used as instrumental variables. This will provide more certainty on whether the many negative results were just caused by low power or by the absence of a true causal association.

Fine-mapping and gene prioritization

Linkage disequilibrium between SNPs is both a blessing and a curse for GWAS. On one hand, it makes it possible to probe only a subset of all genetic variants yet still detect associations for a much larger set, either through tagging of non-genotyped SNPs by genotyped SNPs or through LD based imputation to sequenced reference samples. On the other hand, it means that a detected association doesn't necessarily imply a causal role for the associated SNPs. Fine-mapping attempts to identify which ones out of a number of associated SNPs in a LD region have a causal role, and which ones are merely associated because they are in LD with causal SNPs, but this can be complex and costly [94].

To better understand disease aetiology, it may be of interest to identify causal genes, rather than causal SNPs. In many cases the causal gene may simply be the gene closest to the most strongly associated SNP in a region. However, Chromatin Confirmation Capture experiments show that the majority of chromatin loops are formed between regulatory elements and genes which are not directly adjacent to them, suggesting that it could be common for top associated SNPs to lie at a distance from the genes through which their effect is mediated. For example, a SNP associated with body mass index and located in an enhancer region in an intron of the FTO gene has been shown to disrupt binding of ARID5B, which in turn leads to increased expression of the *IRX3* and *IRX5* genes [95]. This means that the local proximity of that SNP to the FTO gene could be a red herring.

Identifying causal SNPs

A range of different approaches have been developed for the fine-mapping of SNPs. Most of them use information on functional annotation of the genome and LD between SNPs, in addition to SNP association statistics on one or more diseases. An algorithm (PICS) utilizing all these kinds of information has recently been applied to 21 autoimmune disorders and identified many putative causal variants by integrating information from different types of functional annotations, including epigenetic marks and gene expression information [96].

Fine-mapping methods can use LD information to either identify causal SNPs within a region that may not have the strongest association signal, but are located in a functional genomic element like an enhancer, or they can use LD information to identify multiple independently associated SNPs, by calculating the association signal conditionally on the association signal of neighbouring SNPs. Traditionally, this would require full genotype data on the trait of interest. However, it has been demonstrated that it is possible to borrow LD information from a reference genotype data set for a conditional analysis, making it possible to apply this approach to traits for which only summary statistics are available ([67]; GCTA-cojo). For this to work well, the LD structure in the reference genotype population should be a good approximation of the LD structure in the population on which the GWAS has been performed.

Fine-mapping can benefit from data on multiple traits. When two traits share regions of significant genetic associations it can be investigated if they share causal loci at those shared regions, or if different loci drive the regional association in each trait. This has been investigated in a Bayesian framework using only summary statistics and resulted in the identification of 341 loci associated with more than one trait across 42 different phenotypes [97]. SNPs associated with two traits form the basis of the previously discussed Mendelian randomization methods. The same study [97] also investigated evidence for causal relationships between the pairs of 42 traits in a bidirectional fashion, where for each pair of traits the evidence for a causal influence of trait X on trait Y was compared to the evidence for a causal influence of trait Y on trait X. The key idea is the same as in traditional Mendelian randomization: If X is one of many causal influences for Y, then an association of a SNP with X should lead to an association of the same SNP with Y. However, the reverse is not true: An association of a SNP with Y would not lead to an association of the same SNP with

X. This approach identified a causal relationship of BMI on triglyceride levels [97] (implemented in the program gwas-pw).

Identifying causal genes

Genome wide association studies suffer from a massive multiple-testing burden, owing to the large number of association tests between SNPs and phenotype. To minimize the number of false positive results, associations are usually required to be significant at a p-value of 5×10^{-8} (Bonferroni correction of a million independent tests). Gene-based tests have a reduced multiple-testing burden (~20,000 independent tests) and give biological meaning to association results. In gene-based tests SNPs are aggregated into larger groups [98] assuming that SNPs exert their effect through nearby genes, which is not always true. The PrediXcan method refines this aggregation step by including external tissue specific eQTL data to predict gene expression levels based on SNP data [99]. This has several advantages over conventional gene-based tests as it limits the multiple-testing burden by only using SNPs which are known to affect gene expression, and the direction of effect of a SNP on expression levels is not lost when aggregating multiple SNPs. By using tissue specific eQTL data, associations can be tested between a phenotype and expression changes in tissues relevant to the phenotype. PrediXcan has been used to identify genes which may play causal roles in amyloid deposition and cognitive changes in Alzheimer's disease [100] and genes associated with Asthma [101]. While PrediXcan requires individual level genotype data, the extension MetaXcan requires only summary statistics and promises similar accuracy, if the right reference population is used for LD estimation [102]. Transcriptome wide association study (TWAS) is a summary statistics based method similar to MetaXcan, which differs in the algorithm used to predict expression [103]. More recently, TWAS has been applied to detect pairs of traits with genetic correlations at the level of predicted expression [104]. A current limiting factor in this and other expression based methods is the quality of tissue specific eQTL data. The previously mentioned methods prioritize genes based on predicted effects of SNPs on expression. This is in contrast to other methods, such as DEPICT, which use gene expression data to predict gene function and prioritize genes at specific loci based on the predicted function [105].

While an association of predicted gene expression and a phenotype is suggestive of a causal role for that gene, pleiotropy is an alternative explanation for this association. That is, the same SNPs could independently lead to expression changes in one gene and via a

different route have an effect on the phenotype. The summary statistics based Mendelian randomization (SMR) method [106] attempts to distinguish between these two scenarios using eQTL SNPs as instrumental variables and gene expression as exposure variable. To guard against spurious results, this method also introduces the HEIDI test, which identifies and excludes regions where multiple linked SNPs are independently associated with gene expression and phenotype. The method was applied to several complex human traits and has identified 126 putatively causal genes. 77 of these genes are not the closest gene to their respective top associated GWAS hit and may have remained undetected in a conventional gene based analysis.

MetaXcan, TWAS and SMR use the same type of data to identify genes of interest. However, there are many subtle differences between the methods which will likely lead to unique results for each method. To date, no systematic comparison of their relative performance has been published.

Examples of applications to psychiatric disorders

PrediXcan has been applied to bipolar disorder, resulting in the identification of two genes, PTPRE and BBX, for which predicted increased expression in whole blood and the anterior cingulate cortex, respectively, was associated with increased risk of bipolar disorder [107]*. SMR has been applied to schizophrenia, highlighting two genes, SNX19 and NMRAL1, with a potentially causal influence* [106]*. The previous two examples have highlighted genes by using eQTL data, but the concept can be extended to account for the fact that chromatin modifications may mediate the association between genetic variants and eQTLs, and that splice-QTLs, rather than eQTLs, may underlie the genetic effect of a SNP. A TWAS study on schizophrenia has incorporated these ideas and has highlighted 157 genes, many of which were identified due to brain specific splice-QTLs* [108].

Detection of genetic heterogeneity

Most genetic studies are based on the assumption that individuals who exhibit similar symptoms or who have been diagnosed with the same disease are representatives of the same underlying biology defined by a common genetic architecture. Under a polygenic disease architecture, each individual is likely to have a unique combination of risk loci, but with each combination drawn from the pool of risk loci. Genetic heterogeneity occurs when

individuals with the same clinical presentation have risk alleles drawn from independent (or perhaps correlated) sets of risk loci, and the genetic risk profile of one or more subgroups of cases departs from that of the rest. This may arise through misclassification of some cases, or through distinct etiological pathways leading to the same disease [109,110]. The inherent phenotypic heterogeneity within psychiatry makes detection of genetic heterogeneity an appealing goal and identification of distinct pathways holds the promise of shedding light on disease aetiology. Furthermore, a biological basis for disease stratification could lead to more personalized treatments [111].

Although the concept of identifying genetic sub-groups is intuitively appealing simulations suggest it is very difficult to find meaningful genetic groupings if each sub-group has a genetic architecture of a large number of loci with small effects. The large number of combinations of risk loci, their small effect sizes, the uncertainty about the size or even about the presence of genetically heterogeneous groups, and the challenging disentanglement from population stratification all contribute to the difficulty of this problem. As a result, even data sets comprising hundreds of thousands of individuals may not provide sufficient power for naïve approaches to detecting even the simplest scenarios of genetic heterogeneity ([112], Maier et al. unpublished results).

The BUHMBOX method [112] frames the question of disease heterogeneity in a different way and sets out to test if two diseases that share a genetic basis ($r_G > 0$) are correlated because the shared genetic risk factors are present in the whole sample (pleiotropy) or are confined to only a subgroup of individuals (identifiable genetic sub-type or potentially due to misdiagnosis) [112]. The BUHMBOX method investigates LD-independent risk loci for disease B in individuals diagnosed with disease A. If only a subgroup of individuals has a higher genetic risk for disease B, this will induce a correlation among the disease B risk loci, which would support the presence of genetic heterogeneity and not pleiotropy. This approach can demonstrate presence of a genetic sub-group without identifying which specific individuals diagnosed with disease A are genetically more similar to those with disease B. The method found evidence for heterogeneity among seronegative rheumatoid arthritis cases, suggesting that they may contain a significant proportion of seropositive cases. The power of the BUHMBOX method depends on the number of cases, number of markers, risk allele frequency, odds ratio and heterogeneity proportion. For example, with

2000 cases and 2000 controls, a heterogeneity proportion of 0.2 and 50 risk loci, the power to detect heterogeneity at a significance threshold of 0.05 is 92% [112].

Alternatively, genetic heterogeneity can be studied by first grouping individuals (for example disease cases) based on non-genetic data and then testing for genetic heterogeneity between these disease subtypes. A recent method follows this approach by jointly modelling the probability for each SNP of whether its frequency differentiates cases and controls and/or differentiates disease subgroups. Applied to type 1 diabetes, this method suggests that cases with and without autoantibodies exhibit a different genetic architecture for type 1 diabetes disease risk [113].

Examples of applications to psychiatric disorders

The BUHMBOX method was used to investigate the shared genetic basis between major depressive disorder and schizophrenia, and found no evidence that suggested that a subset of major depressive disorder cases was genetically more similar to schizophrenia cases, implying that the genetic correlation estimated between the disorders reflect pleiotropy [112]*. Application of this method will become more interesting as sample sizes increase.*

Table 1: An overview of polygenic methods

Program/Method	Data needed	URL
Estimation of h^2, r_G		
GCTA (GREML)	individual-level genotype data	http://cnsgenomics.com/software/gcta/
BOLT-REML / BOLT-LMM	individual-level genotype data	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
LD score regression	summary statistics	https://github.com/bulik/ldsc; http://ldsc.broadinstitute.org/
Polygenic risk prediction		
PLINK	summary statistics + individual-level data	https://www.cog-genomics.org/plink2
PRSice	summary statistics + individual-level data	http://prsice.info/
GCTA, MTG2 (GBLUP, MTGBLUP)	individual-level genotype data	http://cnsgenomics.com/software/gcta/; https://sites.google.com/site/honglee0707/mtg2
BayesR	individual-level genotype data	https://github.com/syntheke/bayesR
LDpred	summary statistics + individual-level data	https://github.com/bvilhjal/ldpred
Causality of phenotypes		
Mendelian Randomization	individual-level genotype data or summary statistics	http://www.mrbase.org/
gwas-pw	summary statistics	https://github.com/joepickrell/gwas-pw
Causality of genes (Gene prioritization)		
gwas-pw	summary statistics	https://github.com/joepickrell/gwas-pw
SMR	summary statistics + eQTL	http://cnsgenomics.com/software/smr/
PrediXcan	individual-level genotype data + eQTL	https://github.com/hakyimlab/PrediXcan
metaXcan	summary statistics + eQTL	https://github.com/hakyimlab/MetaXcan
TWAS / FUSION	summary statistics + eQTL	http://gusevlab.org/projects/fusion/
DEPICT	summary statistics	https://data.broadinstitute.org/mpg/depict/
Causality of SNPs (Fine-mapping)		
PICS (Fine-mapping)	summary statistics	http://pubs.broadinstitute.org/pubs/finemapping/
GCTA (COJO)	summary statistics	http://cnsgenomics.com/software/gcta/
Detection of genetic heterogeneity		
BUHMBOX	summary statistics + individual-level data	http://software.broadinstitute.org/mpg/buhmbox/
Subtest	summary statistics	https://github.com/jamesliley/subtest

Conclusions

For many psychiatric disorders, genetic factors explain more variation in disease risk in the population than any other known risk factors, but only recently has it become possible to resolve the overall familial genetic risk into individual risk factors at the DNA level. The evidence is now conclusive that psychiatric disorders, like many other common disease and disorders are highly polygenic underpinned by thousands of genetic loci, each of which contributes a small amount to the overall genetic risk. After a period in which many candidate gene studies have reported association results which failed to replicate [114], the hypothesis-free GWAS approach has established itself as the dominating paradigm to find associated genetic loci. With ever growing sample sizes, more and more SNPs surpass the stringent p-value threshold for almost all investigated traits. However, it is also becoming clear that the bulk of genetic risk factors remains hidden among those loci that do not achieve genome wide significance. Many of the methods presented in this review leverage the large amount of information that is harboured by genetic variants, regardless of whether or not they achieve significance. While the focus of some methods is on individual SNPs or genes, other methods aggregate over a potentially large number of loci to answer questions such as “What is the combined genetic effect of all measurable SNPs on phenotypic variance?”, “Do these traits have a shared genetic aetiology?” or “Do these traits causally influence one another?”. One thing that all of these methods have in common is that their utility crucially depends on the power to detect an association, which in turn depends on sample size. Larger sample sizes lead to a higher computational burden, but for most analytical questions which have been presented here, there are methods which can utilize summary statistics and thus drastically reduce runtime and memory requirements.

The literature on new methods is ever-growing and while we have tried to present an overview of key methods to help navigation of this complex field, but it is difficult to be fully exhaustive. We have illustrated the methods with some example applications, however we expect the full potential of the data will only be revealed in coming years when studies with half a million or more people will be widely available. A key issue for the field is to develop cost-effective strategies to capture larger sample sizes with both DNA samples and phenotypic data as these are needed to evaluate the extent to which genetic data can explain phenotypic heterogeneity and to fulfil the potential of more personalized medicine.

Financial support

This research is funded by the Australian National Health and Medical Research Council (N.R.W., grant numbers 1078901, 1087889), (P.M.V., grant number 1078037).

Ethical standards

Not applicable.

2

Chapter 2: Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability

Chapter has been published in Current Epidemiology Reports

Chapter 2: Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability

Naomi R Wray[1] & Robert M Maier[1]

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.

Abstract

The genetic basis of complex genetic disease can be quantified by heritability, which is an estimate of the relative importance of genetic to non-genetic factors in contributing to differences between individuals for any given trait. Heritability is estimated from phenotypic records in data sets of families and represents contributions from genetic variants across the frequency spectrum and genetic variants of any kind and function. Advances in technology allow direct interrogation of some kinds of DNA variants. Specific DNA variants identified in the era of genome-wide association studies explain only a fraction of the heritability estimated from family studies as do less common variants identified through whole exome sequencing. If true effect sizes of risk variants are small studies to date may be underpowered to detect individual risk variants, but the studies may be well-powered to detect the total contribution from common risk variants and this has explained some of the missing heritability. Here we review explanations for the so-called “still-missing heritability” and focus particularly on the issue of genetic heterogeneity.

Introduction

Complex genetic diseases are those that tend to ‘run’ in families yet show no clear pattern of inheritance. Most common diseases are complex genetic diseases including cancers, heart disease, immune disorders and psychiatric disorders. Our understanding of causality of these diseases is limited, and this limited knowledge has contributed to the limited progress made in the development of new treatments. Traditionally, quantification of the genetic basis of disease has been determined by measuring the increased risk of disease in relatives of those affected. Evidence for a genetic risk shared between relatives implies that DNA risk variants are passed from parent to child. This knowledge has underpinned the philosophy that identification of genetic risk variants is a worthy goal that may expose and open new doors towards understanding of causality of disease, which in turn may lead to new treatments. Strategies to identify DNA risk variants have been dictated by available genotyping technologies. Advances in technology of the last decade have delivered methodologies, notably genome-wide association studies (GWAS) and whole exome sequencing (WES) that have started to deliver DNA risk variants associated with disease. Here we review the portfolio of strategies used to understand of the genetic contribution to

complex disease. We close by focussing on the issue of genetic heterogeneity of disease.

Heritability

Evidence for a genetic contribution to disease comes from measurement of an increased risk of the disorder in relatives of those affected. However, such increased risks need to be interpreted with care, since close relatives share a common family environment so that recurrence risk in relatives may also reflect non-genetic factors. Estimates of risks of disease in different types of relatives (e.g. monozygotic and dizygotic twins, first and second degree relatives) are needed to disentangle genetic from non-genetic factors. These risks to relatives are used to estimate heritability on the liability scale [115,116]. Liability to disease is a non-observable or latent, continuous variable with those ranking highest on liability being affected. Heritability on the liability scale, h^2, quantifies the proportion of variance of liability to disease attributable to inherited genetic factors. Comparison of the relative importance of genetic factors for different disorders is more intuitive on this scale, particularly when comparing diseases of different lifetime risk. Heritability accounts for genetic factors that are additive on the liability scale; these genetic factors combine non-additively on the disease scale [117], so that the probability of disease is many times higher for individuals carrying a high number of risk alleles compared to those carrying only half the number. Non-genetic factors include identifiable (but perhaps not recorded) environmental factors or measurement error, but also unidentifiable factors which form an intrinsic stochastic noise. Estimates of heritability may vary between populations, across ages and may depend on whether non-genetic factors have been recorded and included in the analysis [118]. They depend on baseline risk of disease in the population, and the degree of sampling variance is often overlooked. Hence, in reality heritability estimates should be viewed as pragmatic benchmarks representing evidence for low, moderate or high contributions of genetic effects.

Genetic architecture

While heritability on the liability scale expresses the proportion of the variance in liability that is attributable to genetic factors, it tells nothing about the underlying genetic architecture of the disease in terms of number, frequency and effect sizes of individual causal variants, nor

of the mode of action of causal loci (i.e. additive or non-additive). Lack of evidence that complex disease cases represented single gene disorders generated theories of polygenicity [119]. Empirical results of the last decade provide support for a polygenic model [11]. Under a polygenic model, the liability to disease reflects multiple genetic and non-genetic effects acting additively. Hence, liabilities are assumed to be normally distributed, because such a distribution results from many additively acting effects. All individuals in the population carry some genetic risk variants and likely experience some non-genetic risk factors, but most individuals in the population are not affected - disease status results when the cumulative load exceeds a burden of risk threshold.

***De novo* mutations**

De novo mutations are genetic variants present in the DNA of a child but not of their parents. Genotyping of parents and their child is used to identify de novo mutations. Whole exome sequencing has identified that *de novo* mutations play an important role in Mendelian diseases [120]. Effect sizes of *de novo* mutations, that is their contribution to risk of disease, are expected to be both small and large. In contrast, genetic variants of large effect size are more likely to be de novo as they have not been subject to selection. Sequencing studies of the last decade have demonstrated that *de novo* mutations play an important causal role in some complex diseases and disorders for some individuals [121] (for example, mental retardation [122] and autism [123]. For other diseases and diseases there is evidence of an increased burden of de novo mutations in cases compared to controls [124], without being able to identify which of the de novo mutations are individually causal, which increase risk of disease and which are benign [125]. In rare instances, somatic de novo mutations have been shown to be causal [126]. *De novo* mutations are not shared between relatives (except possibly between identical twins, or between siblings as a result of germline mutations in sperm) and so rarely contribute to explaining heritability [127].

Familial vs sporadic

It is not uncommon for cases to be referred to as either ‘familial’ or ‘sporadic’, reflecting whether or not there is a known family history for the disease. In childhood disorders, cases

are similarly referred to as multiplex or simplex depending on presence or absence of other affected children. In common parlance, the terms tend to be interpreted as implying a genetic or non-genetic aetiology of disease, but this can be misleading. On the one hand, knowledge of family history can be used in optimal experimental design. For example, genetic studies designed to identify de novo mutations would be optimised by genotyping of cases with no family history of disease. In contrast, genetic studies designed to identify common genetic risk variants are optimised by prioritising selection of cases with family history and controls with no family history of disease. On the other hand, it is frequently overlooked that under a polygenic genetic architecture the majority of cases are not expected to report family history. For example, for a disease with lifetime prevalence of 1% and heritability of 80% less than a quarter of cases are expected to report family history when considering all first, second and third generation relatives [128]. Likewise, for the same disease more than 60% of monozygotic twins are expected to be discordant for disease status [129].

Missing heritability

Advances in genotyping technology allow cheap genome-wide interrogation of single nucleotide polymorphism (SNPs). GWAS identify associations between SNPs and disease. Reported results from association analyses include risk allele frequency (RAF), effect size (expressed for disease as the odds ratio, OR) and p-value of association. The contribution of these associated DNA genetic variants to variance can be calculated on the liability scale [130] to allow direct comparison of the contribution to risk of each locus on the same scale as heritability is reported. Assuming independence (and ignoring potential overestimation of effect size due to winner's curse), the contribution of each genome-wide significant (GWS) locus can be summed to determine the proportion of variance in liability explained by these loci together, thus quantifying the effects of all genome-wide significant SNPs (h^2_{GWS}).

Given the stringent significance threshold applied, the ability to detect risk loci (i.e., the power) depends on whether the sample size is sufficient given the true effect sizes. When the first GWAS were planned the distribution of expected effect sizes was unknown and sample sizes were powered to detect OR > ~1.3. The first generation of GWAS yielded few GWS results with h^2_{GWS} much less than h^2. This difference has been termed "missing heritability" [7]. As sample sizes have increased, the number of GWS variants have

increased for both quantitative traits and diseases (see Figure 2 in Visscher et al [11]) providing empirical evidence that common variants do play a role in complex genetic disease. None-the-less substantial missing heritability remains.

Hiding Heritability

The observed increase in number of significant association results as sample sizes have been increased [11], implies that the earlier studies were underpowered to detect the variants given their effect sizes. However, given that collection of larger samples is time consuming and expensive, can we be sure that the same will be true for other diseases? Statistical methods that combine quantitative and population genetic concepts to evaluate the contribution to variance of common SNPs across the whole genome without identifying them individually have been developed [18,26,47,131–133]. These methods use people unrelated in the conventional sense of the word, but given the finite global population size, share a proportion of their DNA by descent. The proportion of sharing between pairs of individuals can be estimated using genome-wide marker data, and that genomic similarity can be correlated with disease status to estimate genetic variation [18,21,47,134]. By using distantly related individuals, a significant heritability tagged by common SNPs, h^2_{SNP}, is detected if case-case pairs and control-control pairs have higher genomic similarity than case-control pairs [21]. For most disease traits studied, significant SNP heritabilities have been estimated which demonstrates that although the data sets analysed may have been underpowered to detect the individual small effects as GWS, contributions from common variants exist but that larger sample sizes are needed for individual detection. Hence, the polygenic analyses have been successful in identifying “hidden heritability”, i.e. the increase from h^2_{GWS} to h^2_{SNP}. In theory, with sufficiently large sample size, h^2_{GWS} can become as large as h^2_{SNP}.

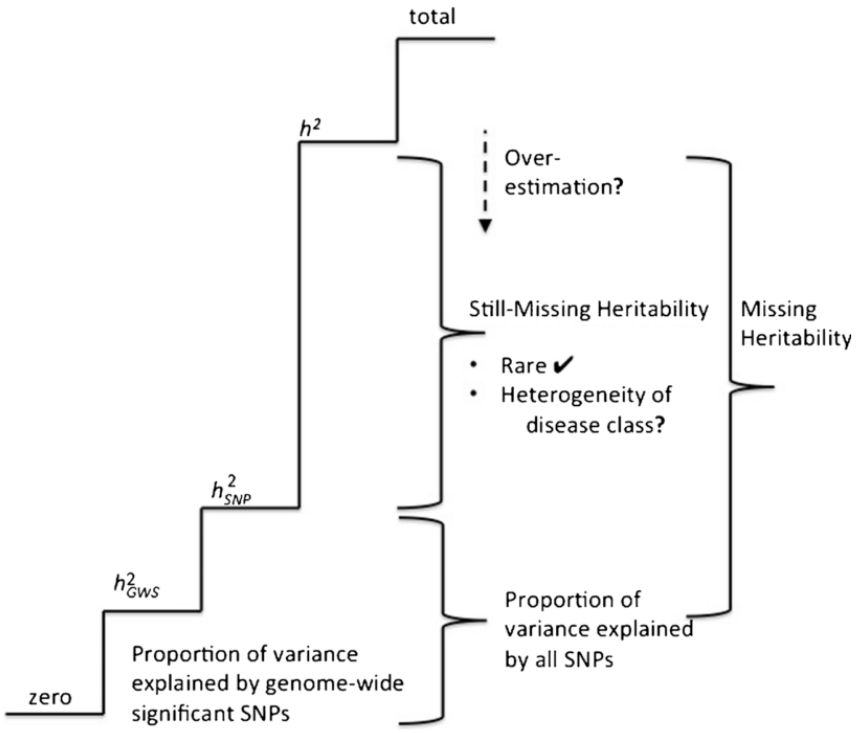

Figure 4: Schematic of heritabilities

Explanations for the still-missing heritability

For most diseases the "still-missing" heritability, i.e. the difference between h^2_{SNP} and h^2 remains substantial at approximately half of the heritability estimated from family data. It is important to note that it is not necessary to explain all heritability when the goal is to open new biological research doors that may impact treatment, and indeed it is likely to be impossible to do so. None-the-less, seeking further insight for the still-missing heritability may also provide important guidance of future research directions. A number of explanations have been proposed [6,7] which include:

a) Over-estimation of heritability from family studies

In human populations, part of the still-missing heritability may simply reflect overestimation of h^2 since typical study designs for estimation of heritability use very close relatives (e.g., full siblings and twins) who share non-additive gene combinations and a common

environment and these confounding factors can be difficult to separate [118,135]. The difference between estimates of h^2 from family data and the "true" h^2 has been termed "phantom heritability" [136] when the difference is attributable to non-additive genetic variance, but our ability to quantify this based on realistically collectable data is limited. Others have argued that the contribution from non-additive genetic variance to complex traits is likely limited [137,138] and that presence of important epistasis and small epistatic variance are not inconsistent [139]. The extent to which gene-environment interaction (GxE) or G and E correlation inflate estimates of heritability from twin and family studies is unknown. Nonetheless, it seems intuitive that exposure to environmental risk factors increases risk of disease only in those that are already genetically susceptible and hence SNP effect sizes may differ in cases stratified by environmental exposure. However, GxE studies to date are limited by a dearth of samples that are informative for G and consistently recorded E [140]. For this reason, studies of candidate GxE interactions have generally lacked replication and the field is plagued by publication bias towards studies with positive results [141].

b) Variants not tagged by common SNPs

Part of the still-missing heritability must reflect genomic variants not well tagged by SNPs [7,18]. Since the SNPs on SNP chips are chosen because both their alleles are common they cannot be in high r^2 linkage disequilibrium with rare causal variants. For many diseases, copy number variants or other rare variants have been identified usually through WES studies. In order to have been detected, necessarily these rare variants have relatively large effect size, but still because they are rare, their contribution to risk in the population is small. A very large number of rare variants are needed to explain the still-missing heritability. For example, a locus with risk allele frequency 0.0001 and heterozygous relative risk (RR) of 10 explains approximately the same proportion of variance in liability as a locus with allele frequency 0.5 and RR 1.06. It is notable, that estimation of h^2_{SNP} using SNPs imputed to the 1000 Genomes reference panel does not tend to generate higher estimates compared to imputation to the HapMap3 panel [142,143]. It is notable that the relative importance of small structural variants to genomic variation is currently not well documented and may not be well represented in sequenced reference panels used for imputation. Since recurrent tandem repeat polymorphisms are known to modulate a range of biological functions [144,145] these may represent an example of an important, but as yet unprobed, source of disease

associated variation. Estimation of h^2_{SNP} based on haplotypes constructed from SNPs is a field of active research, since haplotypes have the opportunity to tag uncommon structural variants not present in imputation reference panels. In practice, such methods may be difficult to apply since they are likely to be very sensitive to genotyping error.

c) Disease heterogeneity

Disease heterogeneity is a possible explanation for still-missing heritability. We have previously noted, for psychiatric disorders at least, that heritabilities estimated from large population samples are lower than those estimated from twin studies. We argued [146] that this may reflect greater diagnostic heterogeneity in large cohorts compared to the carefully collected twin samples, but that the large cohorts may be more representative of the samples currently brought together for analysis in genetic studies.

Disease heterogeneity can have several interpretations, but at its most tangible there are multiple examples of complex genetic diseases that are now recognised to have biologically determined subtypes reflecting independent, or more likely correlated, diseases which may have different optimal treatment strategies. For example, decades ago based on clinical symptoms alone the inflammatory bowel diseases ulcerative colitis and Crohn's Disease would have been indistinguishable and given the same diagnosis. More recently, it has been recognised that diagnosis and treatment of rheumatoid arthritis should consider presence and absence of anti-citrullinated-protein-autoantibodies [147]. The genomics era has allowed good progress in subtyping of cancers (e.g., ER +ve/ER –ve and over-expression of HER2 as a breast-cancer subtype [148,149] or K-ras mutations in colorectal cancer and EGFR mutations in lung cancer, reviewed in [150]), however other branches of medicine are less able to supply measures of phenotypic heterogeneity in the tissue of relevance for mapping onto the genetic heterogeneity. Given the known examples, it seems likely that other diseases currently treated as a single disease entity may in fact be a diagnostic aggregation of sub-types. How could this impact missing heritability? We consider the impact of disease heterogeneity on estimates of the different parameters of variance explained by genetic factors and demonstrate that it could make an important contribution to still-missing heterogeneity.

Exploring the impact of disease heterogeneity

To consider the impact of disease heterogeneity on genetic interpretation of disease, we consider an extreme example of two diseases each of lifetime prevalence 0.5% and heritability 80% that are phenotypically and genetically independent but that have such similar clinical presentation that they are indistinguishable and are considered a single disease. We further assume that both diseases exhibit a genetic architecture in which the effects of all SNPs follow a normal distribution and there is no enrichment of high effect SNPs in high LD regions. The various heritability estimates of the composite disease should therefore not be influenced by LD. Under this composite disease aetiology what would be the impact on estimates of h^2, h^2_{GWS} and h^2_{SNP}?

a) Impact on h^2

The composite disease would have lifetime prevalence of 0.05 (2-0.05) = 0.975% and that the heritability estimated from the two-disease composite would be estimated as greater than 65% from a twin design (see Appendix). In fact, for the composite disease the estimates of heritability using the liability threshold model are expected to be slightly inconsistent when estimated from the relative risks of disease from different types of relatives (**Figure 5**), but such inconsistencies are expected to be difficult to detect given the sampling error on estimates especially since most studies to estimate heritability use relatively small samples of only twins or first degree relatives. We conclude that high estimates of heritability are possible for a composite disease.

Figure 5: Estimates of heritability under a liability threshold model

Calculated from lifetime risk of disease and lifetime risk of disease in relatives of affected individuals for a composite disease that comprises two independent diseases each of lifetime risk 0.5% and heritability 80%.

b) Impact on h^2_{GWS}

We have previously provided theory to estimate power of association studies in the context of misdiagnosis [47] (see Appendix) which is analogous to the scenario here of a disease composite. In **Figure 6** we show the power of an association study to detect risk alleles of a spectrum of frequencies that have effect size under a multiplicative model of heterozygote relative risk 1.15. For a sample of 10,000 cases of a single genetic disease and 10,000 controls we have > 75% power to detect risk alleles of frequencies 0.2-0.8 at genome-wide significance of 5×10^{-8} (line A). However, for our composite disease (for which we expect risk alleles to be associated with only one of the underlying diseases) an association study of 10,000 cases, of which only half are from the disease impacted by the risk allele, is totally underpowered to detect risk alleles (line B). To demonstrate that this reflects the impact of contamination by the phenocopy disease rather than the reduced sample size of the associated disease, we also show the power of an association study of 5,000 cases and 10,000 controls (line C). To consider a range of disease composite scenarios when the proportion disease 2 cases in the disease composite sample is 0%, 5%, 10%, 20% and 50%, the power to detect a disease 1 risk variant of frequency 0.4 and relative risk 1.15 at

the genome-wide significance threshold of $p < 5x10^{-8}$ is 93%, 87%, 79%, 55% and 3% (assuming 10,000 composite disease cases and 10,000 controls and 0.5% lifetime risk of disease 1).

We conclude that disease heterogeneity can severely compromise the power of association studies and hence estimation of h^2_{GWS}.

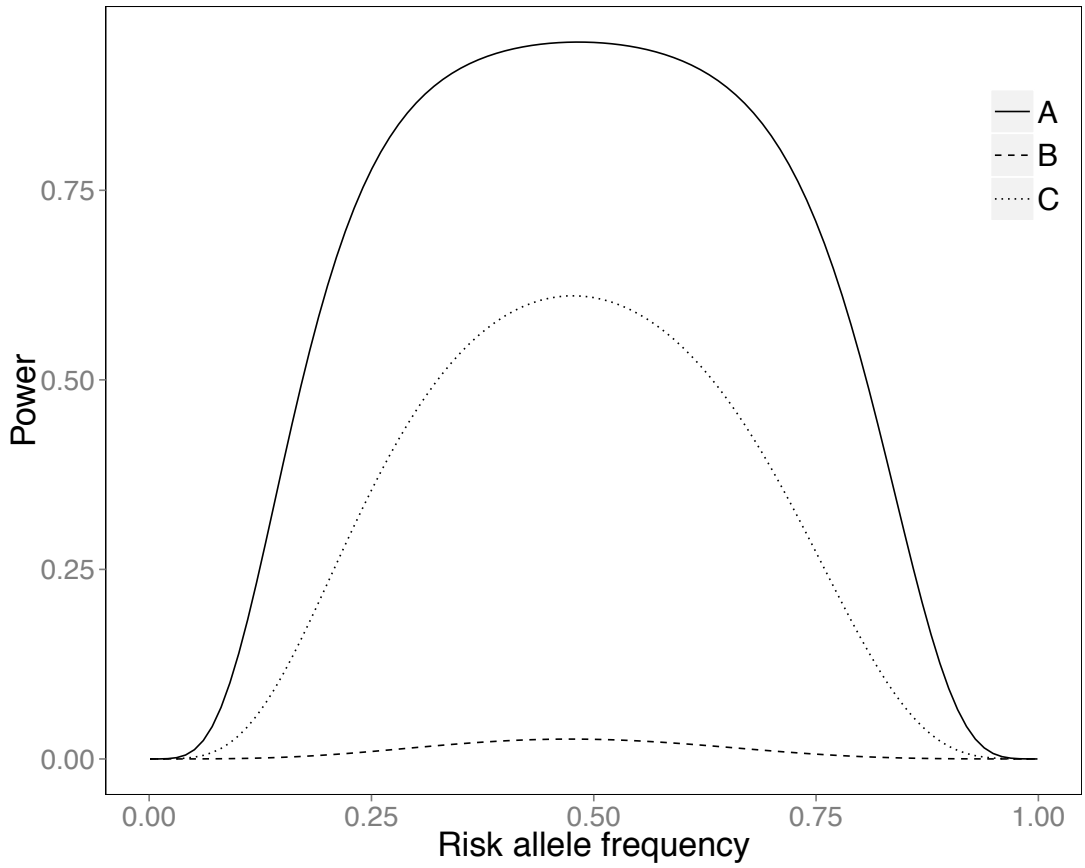

Figure 6: Power of a genome-wide association study to detect risk variants with heterozygous relative risk of 1.15

A) 10,000 cases of a homogeneous genetic disease of prevalence 0.975% and 10,000 screened controls B) 10,000 cases of a composite disease and 10,000 screened controls, the composite disease has prevalence 0.975% but comprises two equally represented genetically independent diseases each of prevalence 0.5% C) 5,000 cases of a homogeneous genetic disease of prevalence 0.5% and 10,000 screened controls

c) Impact on h^2_{SNP}

The impact of analysing a disease composite to estimate h^2_{SNP} can also be considered in terms of disease misclassification [47]. The estimated h^2_{SNP} is a weighted average of the true h^2_{SNP} parameters of each underlying disease and the SNP-covariance (counted twice). So if the two contributing diseases have equal true h^2_{SNP} and are independent the estimated value from the composite disease will be 0.5 h^2_{SNP}. We conclude that disease heterogeneity can generate underestimates of h^2_{SNP} compared to when disease classes are genetically homogeneous.

Summary

The genetic basis of complex genetic disease can be quantified by heritability, which is an estimate of the relative importance of genetic to non-genetic factors in contributing to differences between individuals for any given trait. Heritability is estimated from phenotypic records in data sets of families and represents contributions from genetic variants across the frequency spectrum and genetic variants of any kind and function. Advances in technology allow direct interrogation of some kinds of DNA variants. Specific DNA variants identified in the era of genome-wide association studies, explain only a fraction of the heritability estimated from family studies (h^2_{GWS}) as do less common variants identified through whole exome sequencing. If true effect sizes of risk variants are small then studies to date may be underpowered to detect individual risk variants, but they may be well-powered to detect the total contribution from common risk variants (h^2_{SNP}) and such analysis have helped to explain some of the missing heritability. Here we reviewed explanations for the so-called "still-missing heritability" and focus particularly on the issue of disease heterogeneity. To explore the impact of disease heterogeneity on estimates of h^2, h^2_{GWS} and h^2_{SNP}we considered an extreme example of two independent indistinguishable but equally genetic diseases being lumped together as a disease composite. We have shown that under this scenario the estimates of h^2 from family data are nearly as high as the heritabilities of the contributing individual diseases, yet the estimates of h^2_{GWS} and h^2_{SNP} are severely compromised. In reality this toy example may be too extreme as real presentations of composite diseases may reflect diseases that are genetically correlated rather than totally independent. For example, Crohn's Disease and ulcerative colitis are estimated to have a

genetic correlation based on SNP data of 0.6 [151], which means the vast majority of SNPs identified in GWAS affect both diseases with effects in the same direction, but a handful of them have effects in the opposite direction [152]. Clearly, as the genetic correlation between the two contributing diseases approaches 1, the two diseases merge as a single genetic disease entity. For genetically correlated diseases the power to detect associated loci may be increased by considering the disease composite for loci contributing to both diseases and decreased for other loci. Consideration of these factors can quickly lead to philosophical musings of the definition of disease, since even for a single genetic disease under a polygenic model of disease each individual could carry a unique portfolio of risk loci. In the genomics era, a disease definition may be at the pathway level, whereby a single genetic disease considers different portfolios of risk loci impacting the same pathway, or more practically the class of individuals who respond to the same treatment.

Acknowledgments

NRW is funded by the Australian National Health and Medical Research Council grants 61602 and 1050218.

Appendix

Estimation of heritability from a disease composite

We define a disease composite as a clinically indistinguishable disease comprising two independent diseases. For illustration and simplicity, we assume that the two independent diseases (D_1, D_2) have the same lifetime risk of disease of K and the same heritability of h^2. From standard liability threshold theory, we can calculate the risk in family members whose relatives of a given degree of kinship are affected, K_R. The lifetime risk of the composite disease is $K_C = K(2-K)$. The risk of either of the underlying disease in relatives of those affected by either of the underlying diseases can be written in terms of the probabilities of each of the underlying diseases relatives (D_{R1}, D_{R2})

$$KR\_C = (\mathrm{P}(\mathrm{D_{R1}}\,|\mathrm{D_1}) + \mathrm{P}(\mathrm{D_{R2}}\,|\mathrm{D_1}) - \mathrm{P}(\mathrm{D_{R1}}\ \&\ \mathrm{D_{R2}} \mid \mathrm{D_1}))\,\mathrm{P}(\mathrm{D_1} \mid \mathrm{D_1}\ \text{or}\ \mathrm{D_2}) + (\mathrm{P}(\mathrm{D_{R1}} \mid \mathrm{D_2}) + \mathrm{P}(\mathrm{D_{R2}} \mid \mathrm{D_2}) - \mathrm{P}(\mathrm{D_{R1}}\ \&\mathrm{D_{R2}} \mid \mathrm{D_2}))\,\mathrm{P}(\mathrm{D_2} \mid \mathrm{D_1}\ \text{or}\ \mathrm{D_2}) + (\mathrm{P}(\mathrm{D_{R1}}\,|\mathrm{D_1}\ \&\ \mathrm{D_2}) + \mathrm{P}(\mathrm{D_{R2}} \mid \mathrm{D_1}\ \&\mathrm{D_2}) - \mathrm{P}(\mathrm{D_{R1}}\ \&\mathrm{D_{R2}}|\mathrm{D_1}\ \&\ \mathrm{D_2}))\,\mathrm{P}(\mathrm{D_1}\ \&\ \mathrm{D_2} \mid \mathrm{D_1}\ \text{or}\ \mathrm{D_2}) = \frac{KK_R(2(\frac{1-K}{K} + \frac{(1-K_R)^2}{K_R}) + (2 - K_R))}{2-K}$$

From K_c and K_{R_C}, which are the risks that would be estimable from family data we can calculate the heritability of liability. These calculations have been checked by simulation.

Impact of power of an association study in the context of a disease composite

As before define a disease composite as a clinically indistinguishable disease comprising two independent diseases. A locus is expected to be associated with only one of the two underlying diseases. The underlying disease considered has lifetime risk *K*. We consider a causal variant for this disease that has frequency of the risk allele and protective alleles of *p* and (1-*p*) respectively in the population. Let $(1 - p)^2$, $2p(1 - p)$ and p^2 be the frequencies of the genotypes (in Hardy-Weinberg equilibrium), and the risks of disease in the genotypes are f_0, f_1 and f_2. If we assume a multiplicative model on the disease scale, then $f_1 = f_0\,\gamma$ and $f_2 = f_0\,\gamma^2$ where γ is the relative risk of the risk allele compared to the protective allele. We can calculate the frequency of the risk alleles in cases (true cases) and screened controls as

$p_{case} = \frac{\mathrm{p}\gamma}{1+\mathrm{p}(\gamma\text{-}1)}$ and $p_{control} = \frac{\mathrm{p}}{1-\mathrm{K}}\left(1 - \frac{\mathrm{K}\gamma}{1+\mathrm{p}(\gamma-1)}\right)\frac{\mathrm{p}}{1\text{-}\mathrm{K}}\left(1\text{-}\frac{\mathrm{K}\gamma}{1+\mathrm{p}(\gamma\text{-}1)}\right)$.

If *s* is the proportion of cases in the association sample that that are from the other underlying disease, then the allele frequency in the composite disease sample is

$p_{caseC} = (1\text{-}\,s)\,p_{case} + sp_{control}$

The non-centrality parameter (NCP) of the X^2 test of association is

$\mathrm{NCP} = \frac{N^2(p_{caseC}-p_{control})^2}{Var(\hat{p}_{caseC}-\hat{p}_{control})} = \frac{Nv(1-v)(p_{caseC}-p_{control})^2}{\bar{p}(1-\bar{p})}$

where $\bar{p} = vp_{caseC} + (1-v)p_{control}$ where $v = N_{case}/(N_{case} + N_{control}) = N_{case}/N$. We calculate power as the normal probability p(Z > T), where Z =√NCP and T is the normal deviate corresponding to the type I probability level, i.e., 5×10^{-8} for genome-wide association. When $s = 0$, the power calculation agrees with the genetic power calculator [153].

3

Chapter 3: Genotype based clustering

Unpublished chapter

Chapter 3: Genotype based clustering

Abstract

Medicine is full of examples where a given set of symptoms can be caused by different biological pathways. Diagnostic tests illuminating these pathways are therefore often used to correctly classify the symptom cluster and to find an appropriate treatment. So far, this approach has not been very successful in psychiatry. With the identification of genetic loci contributing to disease risk in psychiatric disorders, attempts have been made to use this information to define genetically defined subtypes. Here we undertake a simulation study to explore the underlying parameters that would be consistent with success of a genetically-based clustering approach.

Introduction

Defining biologically meaningful classifications of psychiatric disorders is a notoriously difficult problem [5]. Diagnostic tests, often based on molecular markers connected to the pathophysiology of a disease, are commonplace in most fields of medicine and usually allow the conclusive and unambiguous categorization of a patient's symptoms. To date, effective, biologically based tests are not available for psychiatric disorders, reflecting our comparatively poor understanding of affective and cognitive brain functions [111]. Other fields of medicine, such as immunology [154] and oncology [155] are full of examples where a molecular test allows distinguishing between disease subtypes which appear homogenous on a clinical level. In many cases this has important consequences for treatment [156]. It has often been suggested that a similar type of genetic heterogeneity may exist in psychiatric disorders, such as Major depressive disorder [109] and Autism [110], and that not recognizing this kind of heterogeneity may limit our understanding, as well as the effective treatment of these disorders [156].

From a quantitative genetics perspective, there is yet another motivation for studying genetic heterogeneity: In most complex disorders the heritability explained by all genetic markers if

well below the heritability estimated from twin or family studies [130]. While there are several possible explanations for this observation, one reason for this so called 'still missing heritability' is the presence of genetic heterogeneity [157]. If it was possible to identify genetically more homogenous groups, the power of genetic studies to detect associated variants would be greatly increased [157].

There have been several studies which have addressed the issue of genetic heterogeneity in psychiatric disorders. Some merely aimed at identifying the presence of genetic heterogeneity in a population [158], while others had the more ambitious goal of identifying genetic subtypes and grouping individuals accordingly [159]. However, while Arnedo et al. [159] reported to find evidence for distinct schizophrenia subtypes, this claim has been heavily disputed on a number of grounds (see PubMed discussion at http://www.ncbi.nlm.nih.gov/pubmed/25219520). While the potential benefit of identifying genetic clusters is large, it is not clear whether clustering methods are powerful enough to achieve this feat.

Here we aim to address this question via a simulation framework. The framework is based on real genotype data and simulated phenotypes, which represent two different subtypes. After de-labeling the subtypes, it is evaluated how well a clustering approach can recover the two subtypes.

Methods

Data

Simulated genotypes lack the complex structure that is often present in real genotype data and that may have a profound effect on the clustering of samples based on their genotypes. We therefore conducted extensive simulations based on real genotypes. For this genotypes from individuals of European ancestry were obtained from the GERA data set (dbGaP number: phs000674.v2.p2). Related individuals were excluded. The genotypes consisted of 585,652 SNPs.

Simulation of phenotypes

For the simulation of genetically correlated phenotypes, between 10 and 1000 SNPs out of the total of 585,652 were chosen to be causal SNPs. Effect sizes for these causal SNPs of the two phenotypes were drawn from a bivariate normal distribution: $\beta \sim N(0, \Sigma)$, where $\Sigma = \begin{pmatrix} h_1^2 & \sigma_g \\ \sigma_g & h_2^2 \end{pmatrix} / M$, and $\sigma_g = r_G \times \sqrt{h_1^2 \times h_2^2}$. The SNP effects were then multiplied with the genotype matrix to obtain an aggregate genetic effect (breeding value) for each individual. Environmental effects were simulated in a similar way for each individual and added to the genetic values, according to the model: $y = g + e$. Two case-control phenotypes were simulated with a prevalence of 0.1, heritability of 0.8 and varying levels of genetic correlation (r_G) between the two diseases. This was done by first simulating normally distributed liability values, and then classifying individuals as cases or controls, depending on whether their liability was above or below the prevalence dependent threshold.

The two phenotypes represent the “true” underlying disorders which should be recovered by the clustering approach. They are then merged into a genetically heterogeneous set where every sample is labeled a case if they were a case in either the first or the second phenotype; otherwise the sample is a control.

Initially 5200 cases and 5200 controls were simulated for each phenotype, which were based on 5000 causal SNPs. To allow for the exploration of a wider parameter space we later switched to another simulation setup which was based on 100 cases and 100 controls for each phenotype, and 10/100/1000 causal SNPs.

Estimation of causal SNPs

Clustering genotypes will inevitably lead to two or more clusters of individuals, but what these clusters represent depends among other things on the SNPs that are used for clustering. If the clusters should capture the two simulated diseases, the clustering must be based on the causal SNPs which were used to generate them, specifically SNPs with different effects on each disease. In any real setting the differentiating causal SNPs are not known and can only be estimated through a GWAS. We therefore performed a GWAS between the heterogeneous cases and the control. Under large sample sizes and low r_G the

most significant SNPs in the GWAS should be the causal SNPs. While this is not necessarily the case anymore with smaller sample size and higher r_G., these SNPs still represent useful candidates for clustering.

Clustering and evaluation

K-means clustering was performed on the genotypes of all cases in the heterogeneous set to obtain two clusters of samples. The clustering was performed only on the genotypes of SNPs which were most significantly associated with disease status in a GWAS of controls versus merged cases. For the k-means clustering, k was set to two, and the Euclidean distance between the genotype vectors of two individuals was used as the measure of dissimilarity.

Finally, the clustered groups are compared to the real underlying phenotypes to test whether the clustering could recover the original groups. This can be done through calculating the AUC (which ideally would be close to 1) or through estimating the genetic correlation between the two clusters (which ideally would recover the simulated genetic correlation).
An illustration of the main steps of the simulation and clustering workflow is shown in **Figure 7**.

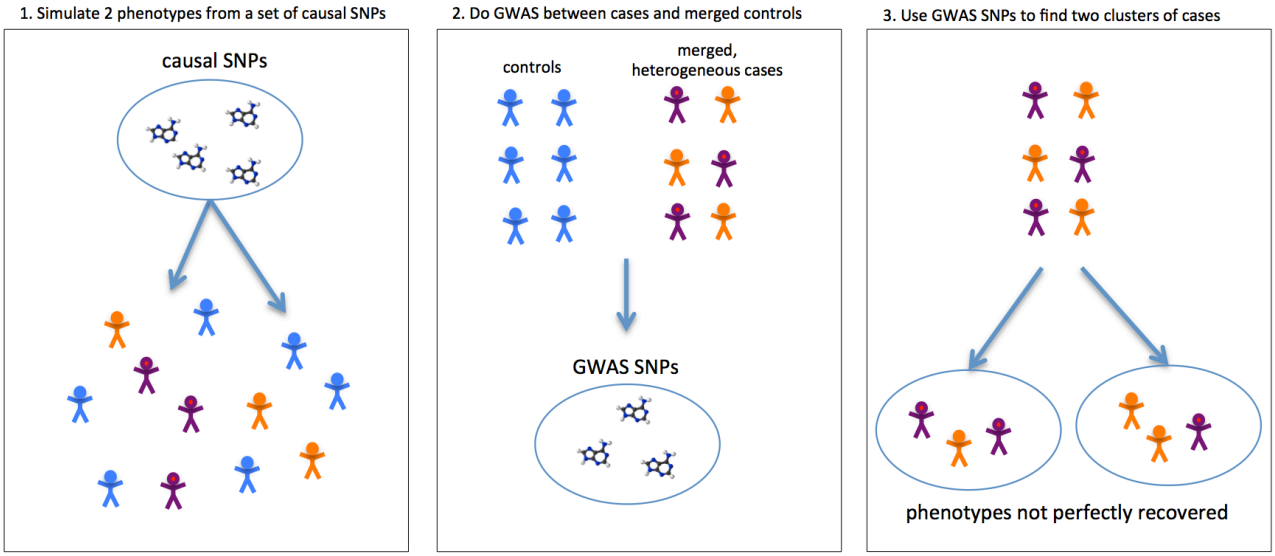

Figure 7: Illustration of the main steps of the clustering analysis

Blue individuals are controls, orange and purple individuals represent genetically heterogeneous but clinically indistinguishable cases.

MDS analysis

The genotype of each individual can be seen as a point in an m-dimensional space, where m is the number of markers. Any type of clustering will only achieve reasonable separation of the two simulated disorders, if the there is a smaller distance between individuals with the same disorder than between individuals with a different disorder. To test whether this is the case, non-metric multi-dimensional-scaling (MDS) was performed on the genotypes of the causal SNPs and of the GWAS SNPs. Similar to PCA, MDS can be used as a dimensionality reduction technique, which maps distances between samples from a high dimensional space to a lower dimensional space, while attempting to keep pairwise distances proportional [160].

Results

In the large-scale simulation setup, the simulated r_G of 0.5 and 0 could not be recovered after clustering the merged case samples using GWAS SNPs (see **Figure 8**).

In order to better understand at which point these simulations fail, a small-scale simulation setup was employed, in which AUC rather than r_G between clusters was evaluated (see **Figure 9**). These analyses suggested that high AUC values are only achieved if the causal SNPs are used for clustering or if the number of causal SNPs is very low.

Impact of simulated r_G on clustering performance.

In order to simulate two different underlying disorders, the SNP effects for the two disorders must have a correlation smaller than 1. The smaller the correlation of the SNP effects, the more genetic differentiation there will be. On the other hand, a GWAS which compares controls to combined cases will be better able to detect causal SNPs if the correlation of SNP effects is high. At a correlation of -1 the SNP effects will be opposite of one another, however if the sample size of the two simulated disorders is equal, a GWAS will not be able to identify the causal SNPs because the effects on disease 1 and disease 2 cancel each other out. Consequently, clustering based on the causal SNPs will work best with a

simulated r_G of -1, but clustering based on GWAS SNPs will work best at higher values of simulated r_G.

The poor performance of the clustering approach raises the following question: Does clustering on genotypes fail to retrieve the original phenotypes, because the clustering algorithm cannot pick up a signal which is present in the genotypes of the causal SNPs or GWAS SNPs? Or does it fail because there is no signal to be picked up at genotypes of the causal or GWAS SNPs? In our simulation the same causal SNPs with different effect sizes give rise to two phenotypes, so it could be expected that the two phenotypes form two clusters of genotypes with different allele frequencies at the causal SNPs. MDS was performed on the genotypes of the causal SNPs and of the GWAS SNPs to answer this question. **Figure 10** shows MDS dimensions 1 and 2 for three simulations with different numbers of causal SNPs. With 10 causal SNPs, the genotypes segregate according to the two different phenotypes. A larger numbers of causal SNPs however does not induce any discernible differences between the genotypes corresponding to the two phenotypes, as can be seen by the nearly perfect overlap of the samples in the MDS figure. In these cases, any clustering algorithm will fail to partition samples according to their simulated disease.

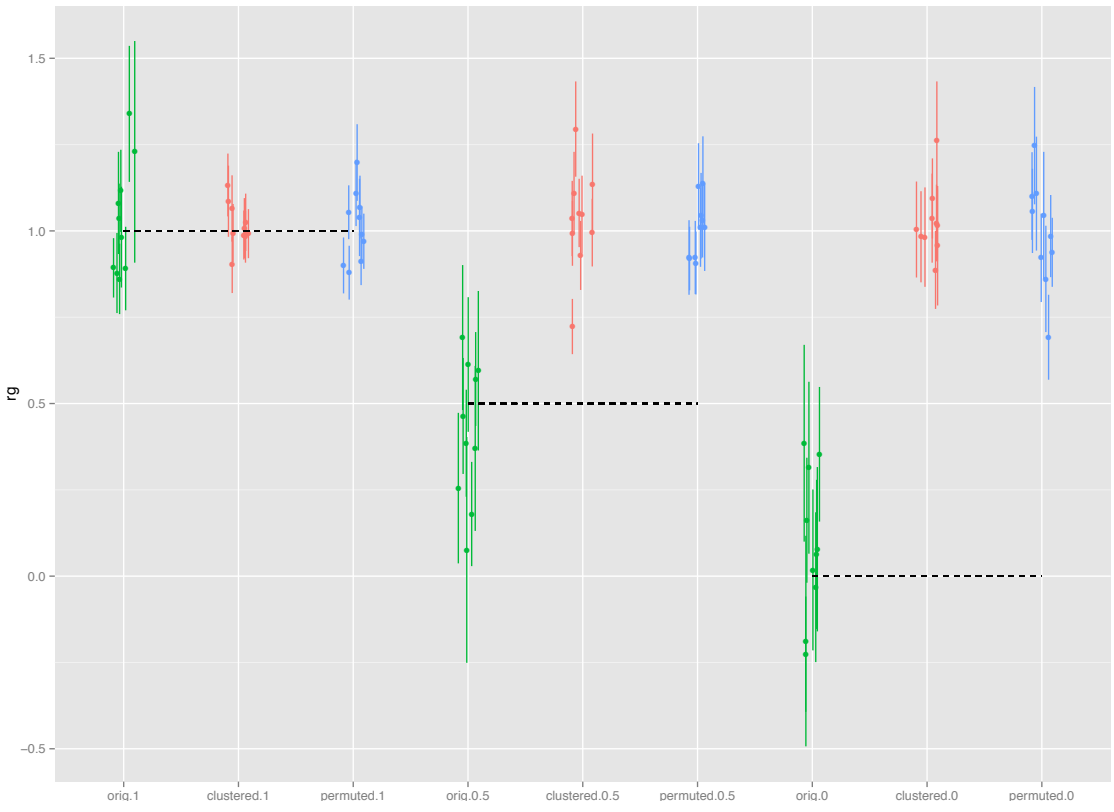

Figure 8: Genetic correlations (r_G) between clusters

The black lines indicate the simulated r_G which would ideally be recovered. The dots and lines represent r_G point estimates and standard errors for each simulation run. Green: r_G not between clusters, but between the original simulated phenotypes to validate the phenotype simulation process. Red: r_G between both clusters. r_G lower than one could not be recovered, indicating that the resulting clusters do not correspond to the simulated phenotypes. Blue: permuted clusters as negative control.

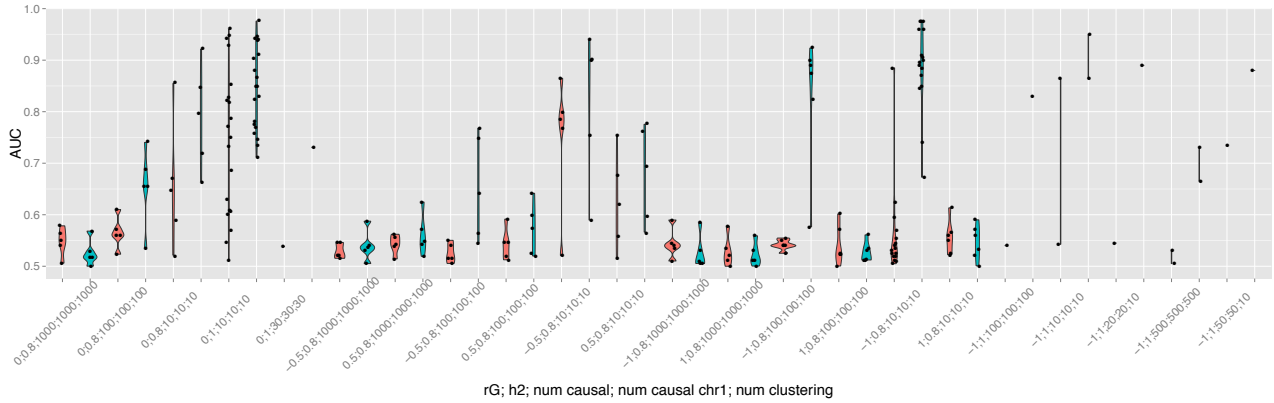

Figure 9: Clustering AUC in the small-scale simulation setup for a wide range of parameters

Pink: Clustering based on GWAS SNPs. Blue: Clustering based on true causal SNPs. Labels on the x-axis show the parameters used for the simulations separated by semicolons: rG, h2, number of causal SNPs, number of causal SNPs which are located on chromosome 1 (always the same as all causal variants here), number of GWAS SNPs which were used for clustering.

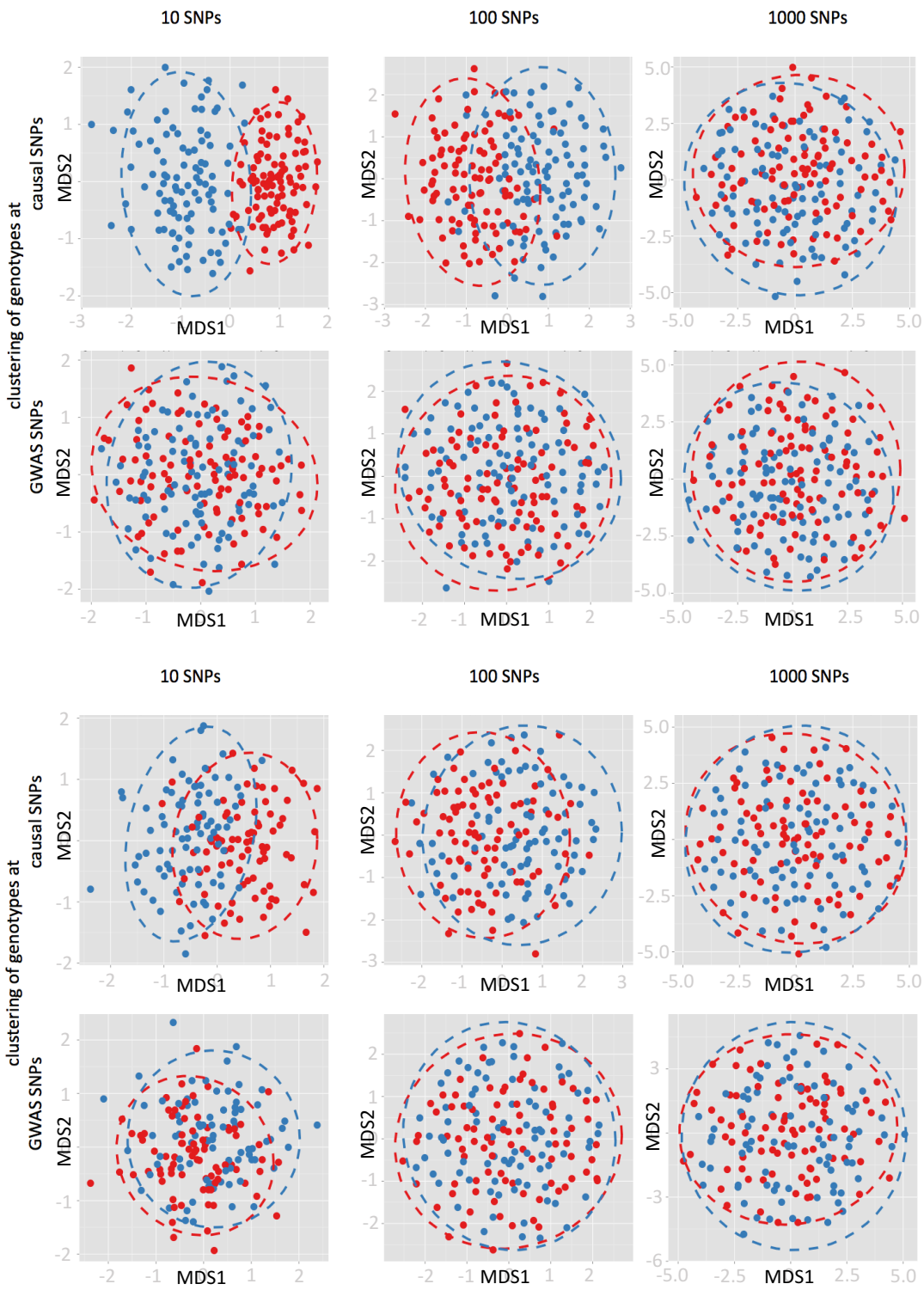

Figure 10: MDS of genotypes performed at causal SNPs and at GWAS SNPs

MDS performed in 6 simulations using 10/100/1000 causal SNPs and r_G of 0. Each dot is a sample which has either disease 1 or disease 2. With small numbers of causal SNPs the genotypes segregate according to the simulated disease. With large numbers of causal

SNPs the two groups become almost inseparable. Top two rows: r_G = -1; bottom two rows: r_G = 0.

Discussion

Current diagnostic classifications in psychiatry are mostly symptom based and may not perfectly align with underlying pathomechanisms [5]. A nosology based on biological mechanisms could greatly benefit diagnostic accuracy as well as enable better targeted treatments [156]. A special case of an imperfect alignment between diagnostic boundaries and causal mechanisms is a scenario in which a clinically homogenous disease consists of two or more groups of genetically different underlying disorders. If the genetic differences between the underlying disorders are large enough, it should be possible to separate the two groups in a clustering approach. In this work, the goal was to study in a simulation setting if, and under which parameters k-means clustering can achieve this task.

The results of our large-scale simulations based on real genotype data did not support the view that clustering can identify groups that correspond to the different phenotypes. We then moved to small-scale simulations based on real genotype data, which allowed us to explore a wider parameter space.

This resulted in the following observations:

(i) Clustering on the true causal SNPs leads to a better performance than clustering on the SNPs which can be detected through a GWAS. This is not surprising, since the SNPs detected in a GWAS are only an approximation to the causal SNPs and effect sizes are estimated with error. Larger sample size and a lower number of SNPs will make this approximation more accurate, as will a higher r_G. However, r_G cannot be too high, otherwise no heterogeneity is being simulated.

(ii) Only if the number of causal SNPs is very small, do the genotypes segregate according to the simulated disease. This is not just a consequence of how well the effects of the causal SNPs can be estimated. The analysis that uses the causal SNPs for clustering demonstrates that even with infinite sample size and perfect estimates of SNP effects, the clustering

approach would fail to identify disease subgroups if the number of causal SNPs is not very small.

Given that the genetic architecture of psychiatric disorders is such that very many small effect loci, but not many large effect loci have been found to contribute to risk [58,109], it appears unlikely that clustering approaches based on genetic data will be very informative, and suggests that claims to the opposite [159] may need to be reconsidered.

It should be noted that our simulations make some assumptions that may not be met in real life applications. Most of the assumptions, if not met, would make it even harder to detect separate groups in genetic data using unsupervised clustering:

First, all simulations assume that the compound disorder consists of two underlying disorders with the same prevalence. In reality, if two underlying disorders are present they are more likely to have different prevalences, which would reduce power.

Second, heterogeneity may not just come in the form of discrete differences. Rather than having two separate underlying disorders with differing genetic effects, there could be a continuum of genetic effects between individuals. In that case, a cluster analysis which is bound to find discrete clusters, would be insufficient to describe the data well.

Third, if two distinct underlying disorders are present, their genetic correlation may be high. Here, many analyses assumed that $r_G = 0$, meaning orthogonal SNP effects for both disorders. In reality this could be much higher, making a genetic differentiation harder. For example, Schizophrenia and Bipolar disorder, despite being clinically distinct, have an r_G of around 0.7 [10].

Being able to stratify a clinical sample into different groups based on their genotypes holds great promises. Not only could it facilitate better targeted treatments, it could also make other studies that depend on biologically well-defined phenotypes more effective. Unfortunately, our simulations indicated that at current sample sizes the power to identify meaningful clusters is very low under a polygenic architecture. It was therefore decided to move on to other projects, such as the summary statistics based multi-trait predictors. Subsequent to our investigations [158] have developed another promising approach to

investigate genetic heterogeneity. Their approach differs from the one presented here. Firstly, because they assume that one of the unobserved underlying disorders includes individuals that are genetically more similar to cases of a second disorder (at the extreme they may be misclassified cases), and its risk loci of the second disorder are known. Secondly, they don't attempt to cluster individuals, but rather to detect the presence of genetic heterogeneity.

4

Chapter 4: Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder

Chapter published in The American Journal of Human Genetics

Chapter 4: Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder

Robert Maier[1], Gerhard Moser[1], Guo-Bo Chen[1], Stephan Ripke[2], Cross disorder Working group of the Psychiatric Genomics Consortium[3], William Coryell[4], James B Potash[4], William A Scheftner[5], Jianxin Shi[6], Myrna M Weissman[7], Christina M Hultman[8], Mikael Landén[8,9], Douglas F Levinson[10], Kenneth S Kendler[11], Jordan W Smoller[12], Naomi R Wray[1] & S Hong Lee[1*]

[1] The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia

[2] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[3]A full list of Cross disorder Working group of the Psychiatric Genomics Consortium may be found in the Supplemental information

[4] Department of Psychiatry, University of Iowa, Iowa City, IA 52242, USA

[5] Department of Psychiatry, Rush University Medical Center, Chicago, IL 60612, USA

[6] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA

[7] Department of Psychiatry, Columbia University, and New York State Psychiatric Institute, New York, NY 10032, USA

[8] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-171 77, Sweden

[9] Institute of neuroscience and physiology, the Sahlgrenska Academy at the Gothenburg University, SE-413 45, Gothenburg, Sweden.

[10] Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA.

[11]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298, USA

[12] Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA

*Correspondence: S. Hong Lee <hong.lee@uq.edu.au>

Abstract

Genetic risk prediction has several potential applications in medical research and clinical practice, and could be used, for example, to stratify a heterogeneous population of patients by their predicted genetic risk. However, for polygenic traits, such as psychiatric disorders, the accuracy of risk prediction is low. Here we use a multivariate linear mixed model and apply multi-trait genomic best linear unbiased prediction for genetic risk prediction. This method exploits correlations between disorders and simultaneously evaluates individual risk for each disorder. We show that the multivariate approach significantly increases the prediction accuracy for schizophrenia, bipolar disorder and major depressive disorder in the discovery as well as in independent validation data sets. By grouping SNPs based on genome annotation and fitting multiple random effects, we show that the prediction accuracy could be further improved. The gain in prediction accuracy of the multivariate approach is equivalent to an increase in sample size of 34% for schizophrenia, 68% for bipolar disorder, and 76% for major depressive disorders using single trait models. Since our approach can be readily applied to any number of GWAS data sets of correlated traits, it is a flexible and powerful tool to maximize prediction accuracy. With current sample size, risk predictors are not useful in a clinical setting but already are a valuable research tool, for example in experimental designs comparing cases with high and low polygenic risk.

Main text

Genome-wide association studies (GWAS) have been highly successful in identifying variants associated with a wide range of complex human diseases [11,161]. However, most common diseases are highly polygenic and each variant explains only a tiny proportion of the genetic variation. Even when associated SNPs are considered jointly in polygenic approaches such as polygenic risk scores [162] or genomic best linear unbiased prediction (GBLUP) [163,164], the accuracy of risk prediction is low. Using more advanced methods [163–167] improved prediction accuracy for traits where a small number of relatively strong associations have been identified, such as type 1 diabetes, ankylosing spondylitis and rheumatoid arthritis, but not for other traits characterized by small effect size variants, including psychiatric disorders [163,164,168].

A major factor determining how well a polygenic model can predict a trait value in an independent sample is the sample size of the discovery data [57,169]. Using more individuals will provide more information and hence increase the accuracy of the estimated effect size of a specific SNP. Sample size can also be effectively increased through data sets measured for correlated traits. Recently, we estimated the genetic relationships between five psychiatric disorders from the Psychiatric Genomics Consortium (PGC) using a bivariate linear mixed model demonstrating that there are significant shared genetic risk factors across the disorders and that measurement of one trait provides information on other genetically correlated traits [10]. Here we extend our bivariate approach to a multivariate linear mixed model and apply multi-trait genomic best linear unbiased prediction (MTGBLUP) [170,171] for genetic risk prediction of disease. MTGBLUP is expected to be more powerful as it uses correlations between disorders and jointly evaluates individual risk across disorders. To date, the information from other correlated traits has been little exploited in the context of risk prediction although recently Li et al. [168] applied bivariate ridge regression to two genetically correlated diseases to improve risk prediction.

An important advantage of the MTGBLUP approach is that it does not require multiple phenotypes to be measured on the same individuals and therefore, can be readily applied to any number of existing datasets of genetically related traits. This is particularly beneficial for disease studies that are limited to a single phenotype but typically aim for large sample sizes. Moreover, it is not necessary for the data sets to be genotyped with the same SNP

array as SNPs can be imputed to a common set of SNPs, such as those available from the HapMap or 1000 Genomes reference panel [172,173]. Prediction accuracy can be expected to improve as more data from phenotypes with shared aetiology are utilised.

In this report, we apply the MTGBLUP approach to the cross-disorder PGC GWAS data and show a significant increase in risk prediction accuracy in independent cohorts of schizophrenia, bipolar disorder and major depressive disorder. MTGBLUP increased the discriminant power between the top and bottom 10% of individuals ranked on their risk predictor, implying that this approach may be useful for stratified medicine in a research setting, to develop tailored interventions or treatments for individuals having different risks [111,156,174]. We further demonstrate a relationship between functionally annotated SNPs and increased prediction accuracy of schizophrenia and bipolar disorder.

As the main method, we use a multivariate linear mixed model for the analyses of GWAS data that estimates the total genetic values of individuals directly by utilising genomic relationships based on SNP information. In the model, a vector of phenotypic observations for each trait is written as a linear function of fixed effects, random genetic effects and residuals. For simplicity, we constrain the description to a single component for the random genetic effects, but the model can be readily extended to multiple components of random genetic effects:

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{e}_1 \quad \text{for trait 1}$$

$$\mathbf{y}_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e}_2 \quad \text{for trait 2}$$

$$\vdots$$

$$\mathbf{y}_n = \mathbf{X}_n\mathbf{b}_n + \mathbf{Z}_n\mathbf{g}_n + \mathbf{e}_n \quad \text{for trait n}$$

where **y** is a vector of trait phenotypes, **b** is a vector of fixed effects, **g** is a vector of total genetic value for each individual and **e** are residuals. The random effects (**g** and **e**) are assumed to be normally distributed with mean zero. **X** and **Z** are incidence matrices for the effects **b** and **g**, respectively. Subscript 1,…, n represents trait 1 to trait *n*. The variance covariance matrix is defined as,

$$\mathbf{V} = \begin{bmatrix} \mathbf{ZA}\sigma^2_{g_1}\mathbf{Z}' + \mathbf{I}\sigma^2_{e_1} & \cdots & \mathbf{ZA}\sigma_{g_{1n}}\mathbf{Z}' + \mathbf{I}\sigma_{e_{1n}} \\ \vdots & \ddots & \vdots \\ \mathbf{ZA}\sigma_{g_{n1}}\mathbf{Z}' + \mathbf{I}\sigma_{e_{n1}} & \cdots & \mathbf{ZA}\sigma^2_{g_n}\mathbf{Z}' + \mathbf{I}\sigma^2_{e_n} \end{bmatrix}$$

where **A** is the genomic similarity matrix based on SNP information, and **I** is an identity matrix. The terms, $\sigma^2_{g_i}$ and $\sigma^2_{g_i}$ denote the genetic and residual variance of trait *i*, respectively and $\sigma_{g_{ij}}$ and $\sigma_{g_{ij}}$ the genetic and residual covariance of trait *i* and *j*. Multi-trait genomic residual maximum likelihood (MTGREML) estimates (see Appendix A) are obtained using the average information algorithm [26,175,176].

Next we show that SNP risk predictors can be easily transformed from individual risk predictors with a simplified BLUP model that uses individual risk predictors as the dependent variable and fits a covariance structure without residual variance (i.e. heritability is 1). Individual risk predictors are the Best Linear Unbiased Predictors (BLUPs) of total genetic value of individual subjects contributed by genome-wide SNPs, i.e. **g** in the previous section. Analogously, SNP risk predictors are defined as the BLUPs of SNP effects estimated jointly with a linear mixed model that intrinsically accounts for linkage disequilibrium between SNPs. The SNP BLUP model is computationally more demanding for a large number of SNPs. Therefore, it is desirable to estimate genetic values (GBLUP) for efficiency, and to transform them to SNP-BLUP. The SNP-BLUP can be projected to predict genetic risk for independent validation sample without the need to have access to the training individuals. The SNP-BLUP estimates can be applied to independent data sets as the SNP weights used to create a risk profile score, for example using the PLINK --score command. The individual BLUP model is

$$\begin{bmatrix}\mathbf{g}_1\\ \vdots\\ \mathbf{g}_n\end{bmatrix} = \begin{bmatrix}\sigma^2_{g_i} & \cdots & \sigma_{g_{1,n}}\\ \vdots & \ddots & \vdots\\ \sigma_{g_{n,1}} & \cdots & \sigma^2_{g_n}\end{bmatrix} \otimes \mathbf{A} \begin{bmatrix}\mathbf{Z}_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \mathbf{Z}_n\end{bmatrix}' \mathbf{V^{-1}} \begin{bmatrix}\mathbf{y}_1 - \mathbf{X}_1 b_1\\ \vdots\\ \mathbf{y}_n - \mathbf{X}_n b_n\end{bmatrix} \qquad (1)$$

SNP BLUP model is

$$\begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_n\end{bmatrix} = \begin{bmatrix}\sigma^2_{u_i} & \cdots & \sigma_{u_{1,n}}\\ \vdots & \ddots & \vdots\\ \sigma_{u_{n,1}} & \cdots & \sigma^2_{u_n}\end{bmatrix} \otimes \mathbf{I} \begin{bmatrix}\mathbf{W}_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \mathbf{W}_n\end{bmatrix}' \Omega^{\mathbf{-1}} \begin{bmatrix}\mathbf{y}_1 - \mathbf{X}_1 b_1\\ \vdots\\ \mathbf{y}_n - \mathbf{X}_n b_n\end{bmatrix}$$

where $\mathbf{W}_i$ is a N × M matrix of standardised SNP coefficients with N being the number of individuals and M the number of SNPs, ⊗ is the Kronecker product function, and the variance covariance matrix for SNP BLUP mode is defined as

$$\Omega = \begin{bmatrix} \mathbf{WI}\sigma_{u_1}^2\mathbf{W}' + \mathbf{I}\sigma_{e_1}^2 & \cdots & \mathbf{WI}\sigma_{u_{1,n}}\mathbf{W}' + \mathbf{I}\sigma_{e_{1,n}}^2 \\ \vdots & \ddots & \vdots \\ \mathbf{WI}\sigma_{u_{n,1}}\mathbf{W}' + \mathbf{I}\sigma_{e_{n,1}}^2 & \cdots & \mathbf{WI}\sigma_{u_n}^2\mathbf{W}' + \mathbf{I}\sigma_{e_n}^2 \end{bmatrix}$$

Replacing **y** with **g** (individual BLUP) and setting residual (co)variances as zero (because individual BLUP is already adjusted for residuals), the variance covariance matrix can be simplified as

$$\Omega = \begin{bmatrix} \sigma_{u_1}^2 & \cdots & \sigma_{u_{1,n}} \\ \vdots & \ddots & \vdots \\ \sigma_{u_{n,1}} & \cdots & \sigma_{u_n}^2 \end{bmatrix} \otimes \mathbf{WW}' = \begin{bmatrix} \sigma_{u_1}^2 & \cdots & \sigma_{u_{1,n}} \\ \vdots & \ddots & \vdots \\ \sigma_{u_{n,1}} & \cdots & \sigma_{u_n}^2 \end{bmatrix} \otimes \mathbf{A}\, M$$

Therefore, SNP BLUP can be written as

$$\begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix} \otimes \mathbf{A^{-1}} \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} M^{-1} \qquad (2)$$

And, this can be rewritten as

$$\begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix}$$

This agrees with Hayes et al. (2009) [177] and Yang et al. (2011) [26] when it reduces to a univariate model. In equation (2), replacing [$\mathbf{g_1}$, … , $\mathbf{g_n}$]' with the right hand side in equation (1), it can be rewritten as

$$\begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix}' \begin{bmatrix} \sigma_{g_i}^2 & \cdots & \sigma_{g_{1,n}} \\ \vdots & \ddots & \vdots \\ \sigma_{g_{n,1}} & \cdots & \sigma_{g_n}^2 \end{bmatrix} \otimes \mathbf{I} \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_n \end{bmatrix}' \mathbf{V^{-1}} \begin{bmatrix} \mathbf{y}_1 - \mathbf{X}_1 b_1 \\ \vdots \\ \mathbf{y}_n - \mathbf{X}_n b_n \end{bmatrix} M^{-1} \quad (3)$$

This agrees with VanRaden (2008) [178] and Stranden and Garrick (2009) [179] derived from a matrix inversion theory when it reduces to a univariate model.

We extended our approach to genomic partitions according to gene annotation. An enrichment analysis based on gene annotation categories has shown that SNPs located within genes identified as being differentially expressed in the central nervous system (CNS) explain a significantly larger proportion of phenotypic variance than expected by chance for schizophrenia and bipolar disorder [10,180]. It is of interest to determine if the

gene/functional annotation information can further increase the prediction accuracy. In the annotation analysis, we grouped SNPs that were located within ±50 kb from the 5′ and 3′ UTRs of 2725 genes differentially expressed in the CNS [19,180] together, and 21% of the SNPs belonged to this category. We then estimated SNP effects from a two component model fitting relationship matrices of SNPs in CNS genes and SNPs localised elsewhere. The model is,

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{g}_{1_{CNS}} + \mathbf{Z}_1\mathbf{g}_{1_{non-CNS}} + \mathbf{e}_1 \quad \text{for trait 1}$$

$$\vdots$$

$$\mathbf{y}_n = \mathbf{X}_n\mathbf{b}_n + \mathbf{Z}_n\mathbf{g}_{n_{CNS}} + \mathbf{Z}_n\mathbf{g}_{n_{non-CNS}} + \mathbf{e}_n \quad \text{for trait n}$$

where g_{CNS} is a vector of random genetic effects due to the CNS genes and $g_{non\text{-}CNS}$ is a vector of random genetic effects due to the non-CNS region.

We also tested another gene set that included candidate genes set for schizophrenia / autism / intellectual disability [162]. We matched these candidate genes with human genome version 18 (on which the discovery data set was built) and retained 4133 autosomal genes. It is noted that we excluded 479 genes flanking GWAS SNPs identified in the Swedish sample [143] to avoid artefact inflation in prediction accuracy. We annotated SNPs within the schizophrenia / autism / intellectual disability genes (28% of the SNPs) and fitted genomic similarity matrices of the annotated SNPs and the rest of SNPs in a two component model.

We had access to the PGC-Cross-Disorder data and three independent validation data sets. The details of the PGC-Cross-Disorder data with additionally available ADHD samples are described elsewhere [10]. Genotype data from each study cohort were processed through the stringent PGC pipeline and imputation of autosomal SNPs was carried out with the HapMap3 reference sample [181]. In each imputation cohort, we retained only SNPs with MAF >0.01 and imputation R^2 >0.6. The number of SNPs used in this study was 745,705. We excluded individuals to ensure that all samples from the 5 disorders were completely unrelated in the conventional sense, so that no pair of individuals had a genome-wide similarity relationship greater than 0.05. The number of cases and controls used in this study are shown in **Table 2**. All phenotypes were controlled for cohort, sex and the first 20 principal components estimated from genome-wide SNPs. Adjustments were performed for each trait.

Table 2: Estimates of SNP-heritability and genetic correlations from multivariate analysis of five psychiatric disorders

Disorders	Cases	Controls	SNP-h^2 on the liability scale	SE
SCZ	8826	6106	0.235	0.011
BIP	5867	3328	0.218	0.017
MDD	8770	6506	0.286	0.023
ASD	3086	3163	0.130	0.024
ADHD	3997	8479	0.281	0.022
			Genetic correlation	**SE**
BIP/SCZ	5867/8826	3328/6106	0.590	0.048
MDD/SCZ	8770/8826	6506/6106	0.365	0.047
MDD/BIP	8770/5867	6506/3328	0.371	0.060
ASD/SCZ	3086/8826	3163/6106	0.194	0.071
ASD/BIP	3086/5867	3163/3328	0.084	0.089
ASD/MDD	3086/8770	3163/6506	0.054	0.089
ADHD/SCZ	3997/8826	8479/6106	0.055	0.046
ADHD/BIP	3997/5867	8479/3328	0.160	0.059
ADHD/MDD	3997/8770	8479/6506	0.242	0.059
ADHD/ASD	3997/3086	8479/3163	-0.044	0.088

SE Standard error; SCZ schizophrenia; BIP bipolar disorder; MDD major depressive disorder; ASD autism spectrum disorder; ADHD attention deficit disorder

In preliminary analysis, using the multivariate linear mixed model, we estimated genetic variances and genetic correlations between the 5 psychiatric disorders (**Table 2**). The estimates agreed with those reported in the previous study [10] (**Figure 11**, **Figure 12** and **Figure 13**) but were slightly less accurate (larger standard errors) because of the smaller sample size due to excluding genetically related samples across all five disorders rather than across only two traits in the bivariate analyses.

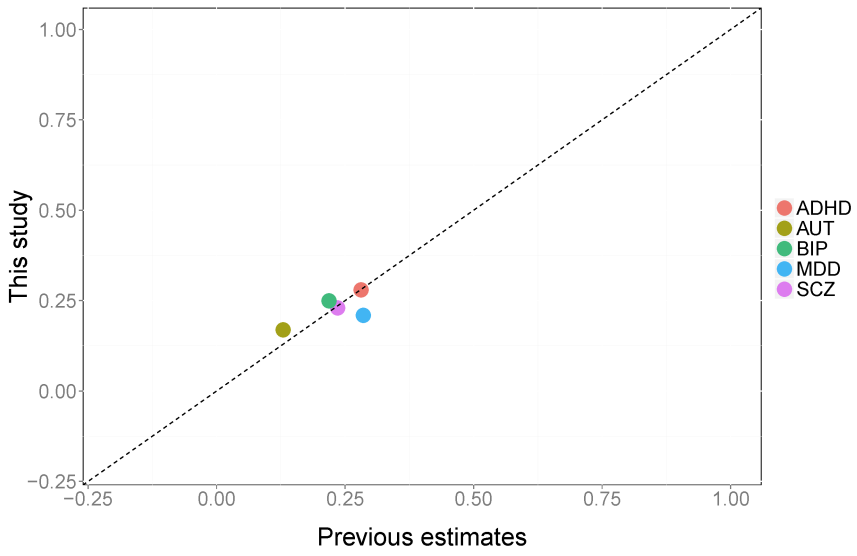

Figure 11: Previous vs current estimates – heritability

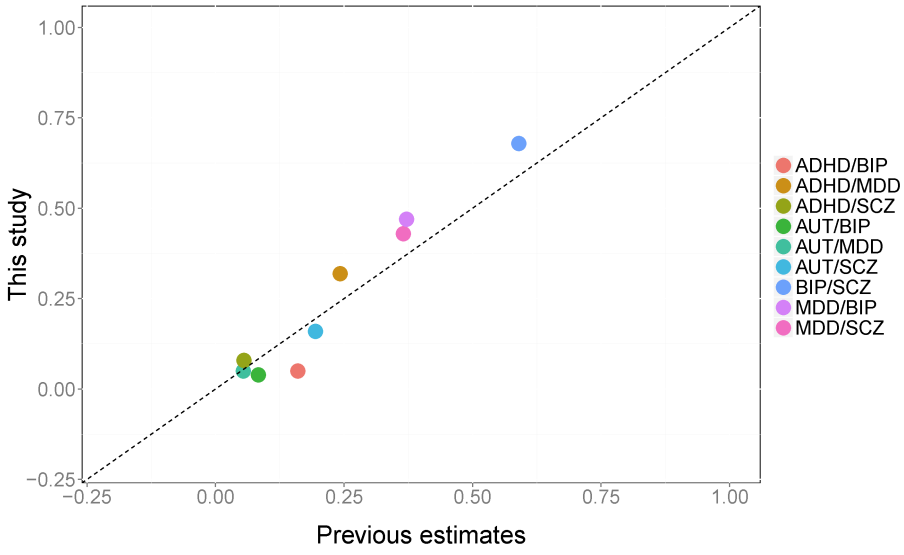

Figure 12: Previous vs current estimates – genetic correlations

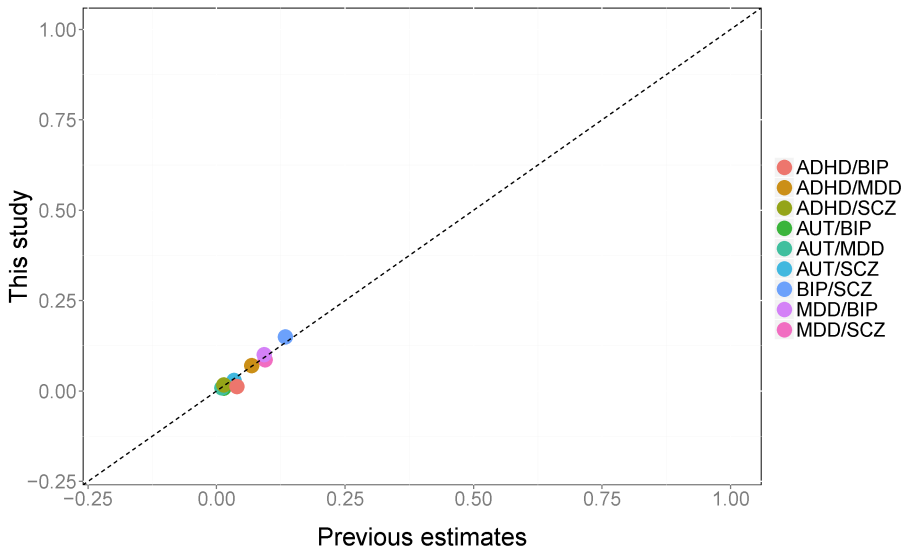

Figure 13: Previous vs current estimates – SNP-coheritability

Previous estimates (Lee et al. 2013) plotted against estimates from this study for 5 psychiatric disorders. Both studies and the previous utilized the same data. However, the previous estimates used a bivariate model and so overlapping and closely related samples were excluded on a pairwise basis. In this study overlapping samples and closely related samples were removed across all 5 disorders generating small samples per disease. ***Figure 11*** *SNP-heritability (correlation coefficient between previous and current estimates = 0.69);* ***Figure 12*** *Genetic correlations (correlation coefficient between previous and current estimates = 0.98);* ***Figure 13*** *SNP-coheritability (correlation coefficient between previous and current estimates = 0.98)*

To evaluate the risk prediction performance of MTGBLUP, we performed within-study cross-validation of the PGC data, i.e. internal validation. We randomly split the data for each disease into a training sample containing ~80% of individuals and a validation sample containing the remaining ~20% [182] and repeated this five times. For assessing predictive performance in the internal validation, we calculated the correlation coefficient between the observed disease status and the predicted genomic risk score of the validation individuals. We also regressed observed disease status was on risk scores. If the risk scores are unbiased estimates of genetic risk then the regression coefficient is expected to be one, i.e. the covariance between true and estimated risks equals the variance of estimated risks. Deviations from one reflect the degree of bias of the risk scores. We averaged the correlation and regression coefficients, and estimated empirical standard errors over 5 replicates. Using

the empirical standard errors estimates, a t-test was performed to assess differences in prediction accuracy between methods. In the within-study cross-validation, MTGBLUP outperformed single-trait genomic best linear unbiased prediction (STGBLUP) for all disorders: the gain in prediction accuracy was significant for schizophrenia (p-value < 6.0E-08) and bipolar disorder (p-value < 6.6E-11) (**Figure 14**). The slope from the regression of disease status on predicted risk score ranged from 0.88 to 1.14 (**Table 3**) indicating that the risk scores are well calibrated.

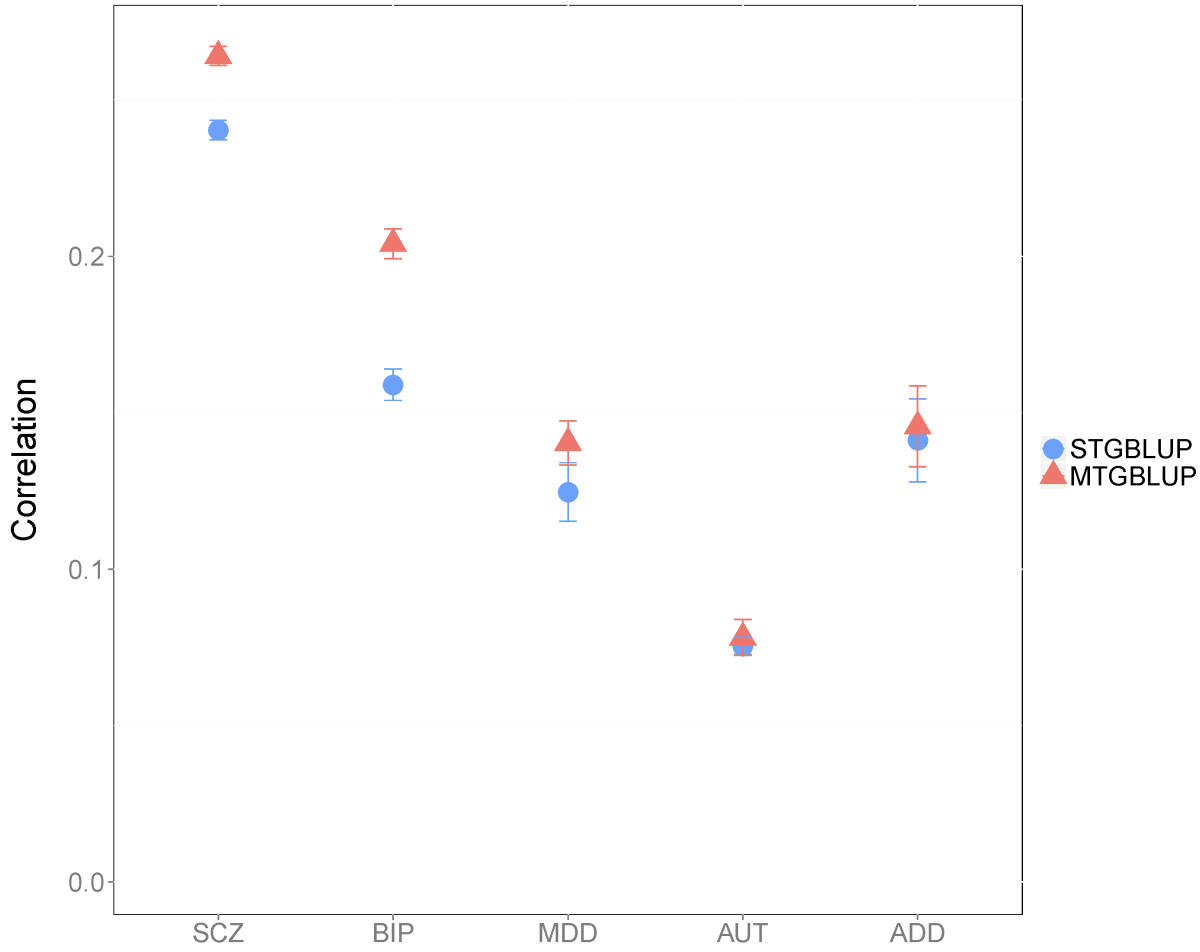

Figure 14: Prediction accuracy of MTGBLUP and STGBLUP for five psychiatric disorders in the within-study validation of PCG

Results are based on 5 replicates. Error bars are ± empirical standard error. Prediction accuracy is measured as the mean of the correlation coefficient between the true disease status and the predicted genomic risk score in the validation data.

Table 3: Comparison of prediction accuracy (correlation) and regression coefficient (Regression) of MTGBLUP and STGBLUP for five psychiatric disorders in the within-study validation of PCG

Correlation and regression coefficient is averaged over 5 replicates. Empirical SE over 5 replicates is in bracket.

	Correlation	Regression
	Schizophrenia	
STGBLUP	0.240 (0.003)	1.011 (0.022)
MTGBLUP	0.264 (0.003)	1.019 (0.019)
	Bipolar disorder	
STGBLUP	0.159 (0.005)	1.091 (0.054)
MTGBLUP	0.204 (0.005)	0.971 (0.025)
	Major depression	
STGBLUP	0.125 (0.009)	1.078 (0.054)
MTGBLUP	0.140 (0.007)	0.930 (0.038)
	Autism Spectrum Disorders	
STGBLUP	0.075 (0.003)	0.965 (0.080)
MTGBLUP	0.078 (0.006)	0.884 (0.054)
	ADHD	
STGBLUP	0.141 (0.013)	1.144 (0.052)
MTGBLUP	0.146 (0.013)	1.116 (0.050)

Results obtained from a within-study validation might not reflect the true performance when SNPs effects estimated from the training data are spuriously associated with the diseases. To better assess the true prediction potential of MTGBLUP, risk scores derived from the complete PCG data were validated in independent samples for schizophrenia, bipolar and major depressive disorder. As independent validation sets, we used Swedish schizophrenia [143] and bipolar GWAS data [183] and the GENRED2 MDD dataset collected by the same methods as reported for the GENRED1 dataset [184]. SNPs in the validation data were processed through the same stringent quality control as the discovery data. The Swedish

schizophrenia data was imputed using the HapMap3 as reference. The bipolar disorder data and MDD data were imputed using the 1000 Genomes Project data as reference. Post-imputation quality control was applied to exclude poorly imputed SNPs from the validation sets. Finally, we selected SNPs that matched those in the discovery set. The number of SNPs in each validation set is shown in **Table 4**. Individuals were removed from the validation datasets if they had relatedness > 0.05 to any one of the individuals in the discovery set. **Table 4** gives the numbers of cases and controls in the independent validation datasets before and after excluding related individuals.

Table 4: Numbers of cases and controls in the independent validation data sets before and after removing related individuals

	SCZ (Swedish)		**BIP (Swedish)**		**MDD (GENRED2)**	
	Cases	Controls	Cases	Controls	Cases	Controls
All	5193	6391	2208	6056	831	474
After cut-off QC	4068	5471	2029	5338	822	466
Number of SNPs	745631		645237		673109	

SCZ: Swedish schizophrenia GWAS, BIP: Swedish bipolar disorder GWAS, MDD: GENRED2 GWAS.

In the discovery set we obtained SNP solutions by applying SNP GBLUP (Eq. (3)) and then projected the SNP solution to the genotypes of the validation individuals (Eq. (2)). For assessing predictive performance in the independent validation, the correlation and regression coefficient were used as measures of prediction accuracy and biasedness, respectively, similar to the internal validation. A likelihood ratio test (LRT) was used to test for differences in prediction accuracy between methods comparing the likelihood of a logistic regression fitting the STGBLUP to that of a logistic regression fitting the MTGBLUP and STGBLUP jointly. In the logistic regression models, case-control status was used as the dependent variable. In the validation datasets, all phenotypes were controlled for cohort, sex and the first 20 principal components just as in the discovery dataset. This external validation confirmed the superior performance of MTGBLUP over STGBLUP (**Table 5**). From the LRT to test differences in prediction accuracy, the model including MTGBLUP fitted the data significantly better (p-value=2.4E-24 for schizophrenia, 6.6E-16 for bipolar disorder and 0.010 for major depressive disorder) (**Table 6**). We further tested a two-components model

fitting similarity matrices based on SNPs annotated in CNS genes and or SNPs localised elsewhere (MTGBLUP-CNS and STGBLUP-CNS). Including the CNS component resulted in increased gain in prediction accuracy for schizophrenia and bipolar disorder (**Table 5** and **Table 6**). We also tested a second annotation model replacing the CNS gene set with a schizophrenia / autism / intellectual disability (SAI) candidate genes set (4133 autosomal genes)[3] (MTGBLUP-SAI or STGBLUP-SAI), but found little improvement due to SAI genes for three of the disorders (**Table 7** and **Table 8**).

Table 5: Prediction accuracy for schizophrenia, bipolar disorder and major depressive disorder in independent validation data sets

Prediction accuracy is given as the correlation coefficient between the observed disease status and the predicted genomic risk score in the validation data. Regression deviated from one reflects the degree of bias of the risk scores.

	Correlation			**Regression slope**		
	SCZ	BIP	MDD	SCZ	BIP	MDD
STGBLUP	0.198	0.129	0.045	0.784	0.709	0.304
MTGBLUP	0.222	0.159	0.075	0.815	0.697	0.466
STGBLUP-CNS	0.203	0.132	0.045	0.789	0.719	0.306
MTGBLUP-CNS	0.224	0.162	0.076	0.807	0.690	0.476

Table 6: P-values from the likelihood ratio test comparing different models

Likelihood ratio LR = -2 [logL(x_1) - logL(x_1+ x_2)] where logL(x_1) (logL(x_1+x_2)) is the log likelihood from a logistic regression with case-control status as the dependent variable and x_1 (x_1 and x_2) as independent explanatory variable.

		SCZ	**BIP**	**MDD**
x_1	x_2	p-values from LRT		
STGBLUP	MTGBLUP	2.4E-24	6.6E-16	1.0E-02
STGBLUP	STGBLUP-CNS	9.1E-06	4.6E-03	5.8E-01
MTGBLUP	MTGBLUP-CNS	2.4E-03	5.3E-03	3.3E-01
STGBLUP	MTGBLUP-CNS	6.7E-26	1.3E-17	7.3E-03

Table 7: Prediction accuracy for schizophrenia, bipolar disorder and major depressive disorder in independent validation data sets when using a second annotation model

MTGBLUP-SAI or STGBLUP-SAI: a second annotation model replacing the CNS gene set with a schizophrenia / autism / intellectual disability (SAI) candidate genes sets. Prediction accuracy is given as the correlation coefficient between the true disease status and the predicted genomic risk score in the validation data.

	Correlation			Regression		
	SCZ	BIP	MDD	SCZ	BIP	MDD
STGBLUP-SAI	0.199	0.130	0.048	0.787	0.746	0.323
MTGBLUP-SAI	0.222	0.160	0.076	0.817	0.718	0.470

Table 8: Comparison of the fit of standard model with the SAI-annotation model for STGBLUP, MTGBLUP and MTGBLUP

Likelihood ratio LR = -2 [logL(x_1) - logL(x_1+ x_2)]

		SCZ	BIP	MDD
x_1	x_2	p-values from LRT		
STGBLUP	STGBLUP-SAI	0.18	0.18	0.070
MTGBLUP	MTGBLUP-SAI	0.22	0.71	0.54
STGBLUP	MTGBLUP-SAI	1.2e-24	9.7E-15	0.0083

When using independent validation samples, the slopes of the regression of the case-control status on the predictor were less than 1 (**Table 5**). The bias was relatively small for schizophrenia and bipolar disorder but larger for major depressive disorder. A slope less than one implies that the difference between the true genetic risks in a pair of individuals is less than that of the predicted genetic risk between them. The bias could be due to low predictive power (e.g. MDD) or to heterogeneity between the discovery and validation sample.

In order to assess population differences, we calculated ancestry principal components from the POPRES reference sample [185,186] and projected them into the discovery and validation samples and found ancestral differences between them for each disorder (**Figure 15**, **Figure 16** and **Figure 17**). We estimated that the SNP-correlation [176] between the discovery and validation data set was significantly different from 1 for schizophrenia and bipolar disorder (**Table 9**; the point estimate was lower for MDD but the small sample size generated a large standard error so it was not significantly different from 1). To explore if the found heterogeneity reflects real population differences or is caused by other factors that lead to differences between the discovery and validation samples such as batch effects, we looked for evidence of heterogeneity within PGC discovery samples for schizophrenia, bipolar disorder and major depressive disorder (Appendix B). For each disorder, we divided the discovery sample into four groups based on the 25%, 50% and 75% quartile of the first principal component, which reflects ancestral population differences between individuals (Figure S4). Applying a reaction norm model [187,188] (Appendix B), we found significant heterogeneity attributable to the ancestral population differences for schizophrenia and bipolar disorder (**Table 10**, **Figure 21**, **Figure 22** and **Figure 23**). This indicates that for schizophrenia and bipolar disorder real population heterogeneity rather than batch effects contribute to the reduced SNP-correlation between discovery and validation set. Previously we reported more heterogeneity between MDD cohorts than between schizophrenia cohorts [10], where cohorts were defined based on sample collection, genotyping platform and imputation set. The lack of evidence of population heterogeneity for the depression sample here may reflect that population heterogeneity not detectable given other heterogeneity within these samples.

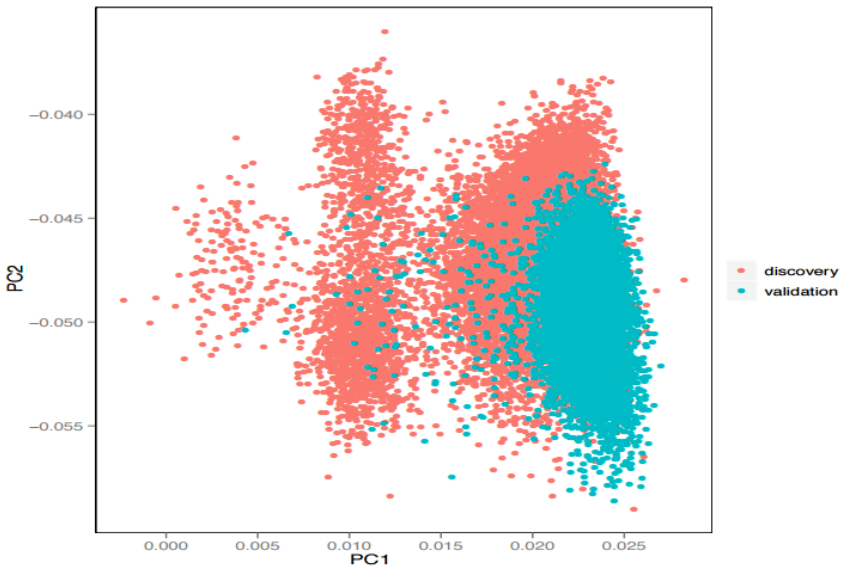

Figure 15: Principal components – schizophrenia

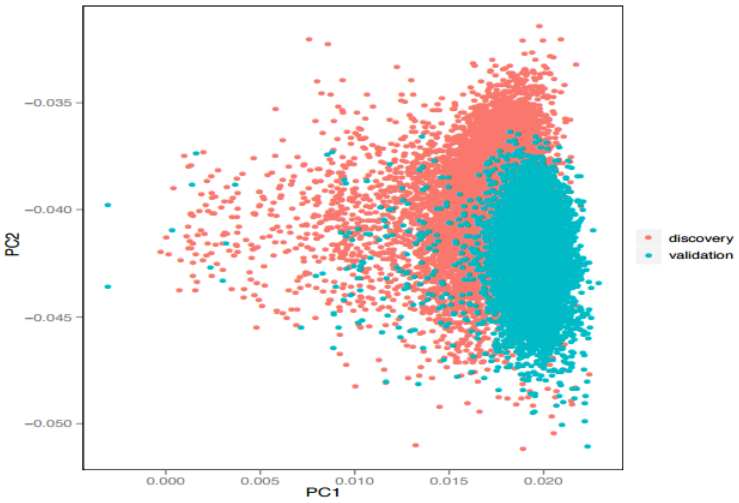

Figure 16: Principal components – bipolar disorder

Figure 17: Principal components – major depressive disorder

*Principal component analysis based on the projected PC from POPRES for SCZ (**Figure 15**), BIP (**Figure 16Figure 12**) and MDD (**Figure 17**). The same SNPs were selected from the discovery and validation set and used to project PC in each disorder. The number of SNPs used was 745,631 for SCZ, 645,237 for BIP and 673,109 for MDD.*

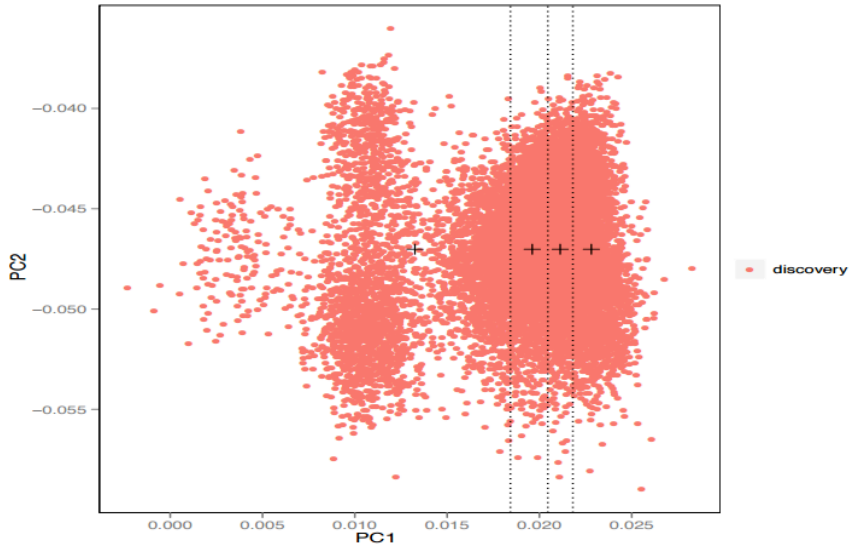

Figure 18: Principal components – quartiles – schizophrenia

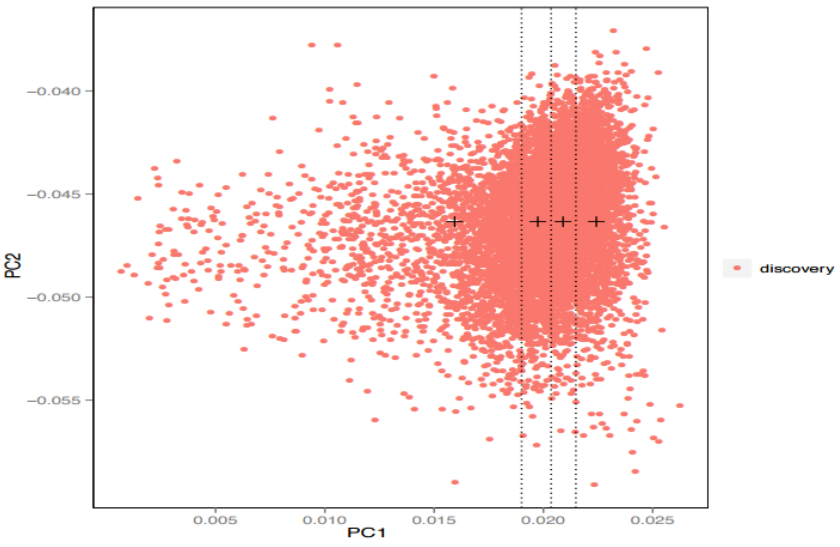

Figure 19: Principal components – quartiles – bipolar disorder

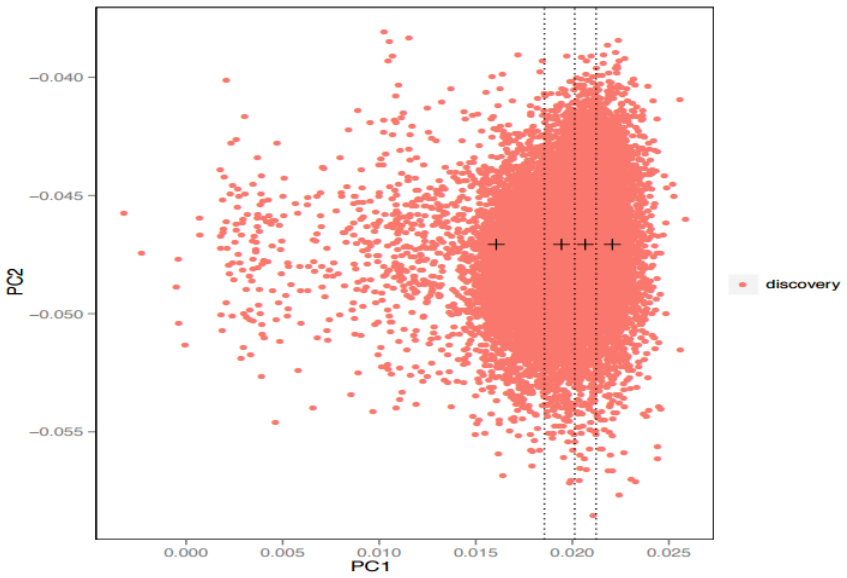

Figure 20: Principal components – quartiles – major depressive disorder

*Principal component analysis based on the projected PC from POPRES for the discovery sample of SCZ (**Figure 18**), BIP (**Figure 19**) and MDD (**Figure 20**). The number of SNPs used to project PC was 745,705 for all three disorders. The dashed lines are 25%, 50% and 75% quartiles of the first principal component in the discovery sample (four population*

classes) and the plus sign is the mean (of PC1) of each population class. The four population classes for each trait were used in the reaction norm model (Appendix B).

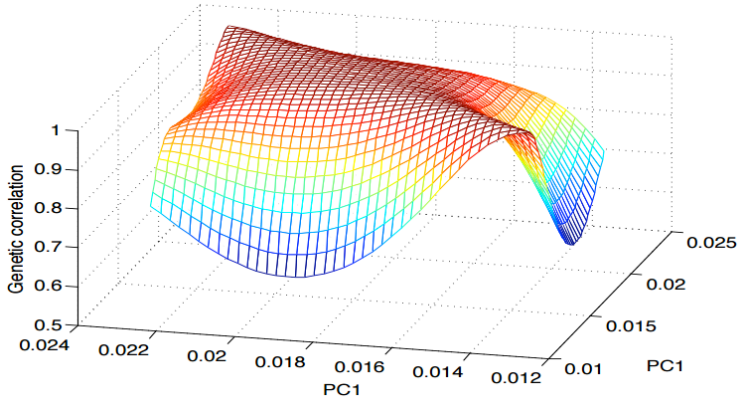

Figure 21: Reaction norm model – schizophrenia

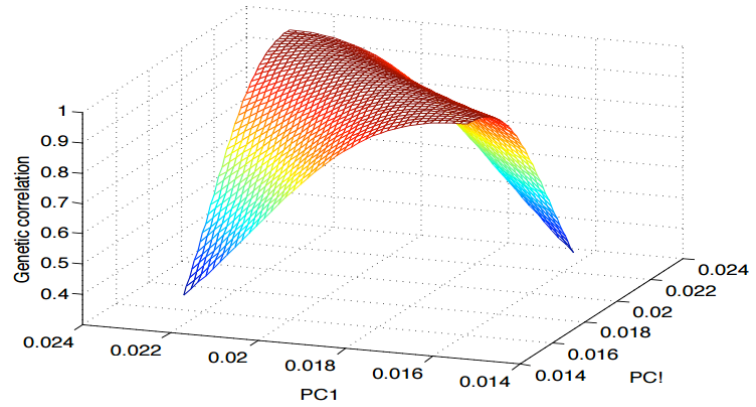

Figure 22: Reaction norm model – bipolar disorder

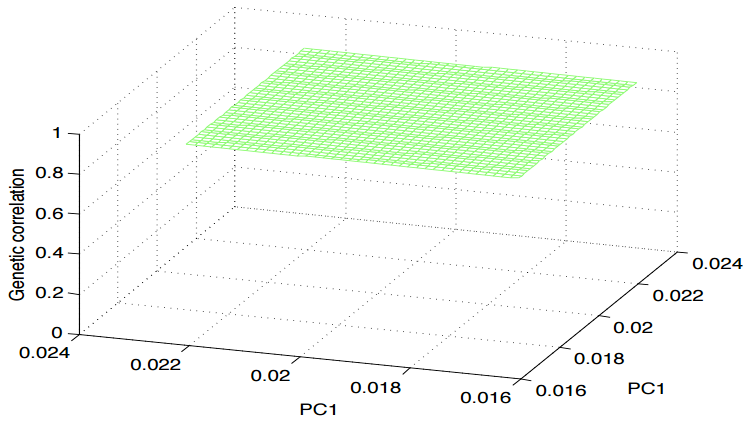

Figure 23: Reaction norm model – major depressive disorder

*Genetic correlation pattern across different ancestry principal components estimated from the reaction norm model (Appendix B). Order of polynomial (see **Table 10**): **Figure 21** k=3 for SCZ, **Figure 22** k=2 for BIP, **Figure 23** k=1 for MDD.*

Table 9: SNP-heritability and genetic correlation from bivariate analyses of the discovery and validation data set for SCZ, BIP and MDD

h^2 is SNP-heritability on the liability scale. r_g is genetic correlation between discovery/validation set. P-value is for testing if r_g is different from 1, indicating heterogeneity for a lower p-value.

Trait 1/ trait 2	Cases T1/T2	Controls T1/T2	Trait 1 h^2 (SE)	Trait 2 h^2 (SE)	r_g (SE)	p-value
SCZ discovery/ SCZ validation	8826/ 4068	6106/ 5471	0.23 (0.01)	0.21 (0.02)	0.80 (0.05)	7.3E-51
BIP discovery/ BIP validation	5867/ 2029	3328/ 5338	0.21 (0.02)	0.22 (0.02)	0.75 (0.08)	7.3E-17
MDD discovery/ MDD validation	8770/ 822	6506/ 467	0.28 (0.02)	0.11 (0.25)	0.51 (0.64)	0.84

Table 10: Reaction norm model to test heterogeneity across populations classified by the first ancestry principal component

Schizophrenia (p-value=0.00078) and bipolar disorder (p-value=0.0017) show a significant evidence for heterogeneity across different populations.

k	log L	Number of parameters	LR	p-value
SCZ				
1	3830.01	5	0.00	1
2	3836.61	7	13.20	0.0014
3	**3840.55**	**10**	**21.07**	**0.00078**
4	3841.36	14	22.69	0.0069
BIP				
1	2342.89	5	0.00	1
2	**2349.27**	**7**	**12.76**	**0.0017**
3	2351.64	10	17.49	0.0037
4	2352.77	14	19.75	0.019
MDD				
1	**3326.17**	**5**	**0.00**	**1**
2	3326.39	7	0.42	0.81
3	3328.24	10	4.14	0.53
4	3330.98	14	9.61	0.38

Following a common epidemiological approach to assess a continuous risk factor [58], individuals were stratified into deciles according to the ranked values of the genetic risk predictors. We estimated the odds ratio of case-control status by contrasting each decile to the lowest decile (**Figure 24**). For all disorders the odds ratio was highest between individuals in the highest and lowest decile, ranging from 1.3 to 5.5. Generally, odd ratios from MTGBLUP were larger than those from STGBLUP. For example, for bipolar disorder MTGBLUP increased the odds ratio by up to 60% compared to STGBLUP (odds ratio of 4.4 and 2.8, respectively). The discriminant power increased more for the annotation model with

the CNS genes, compared to the one-component models without annotation (**Figure 24**). With increasing sample sizes the odds ratio is expected to increase further [58].

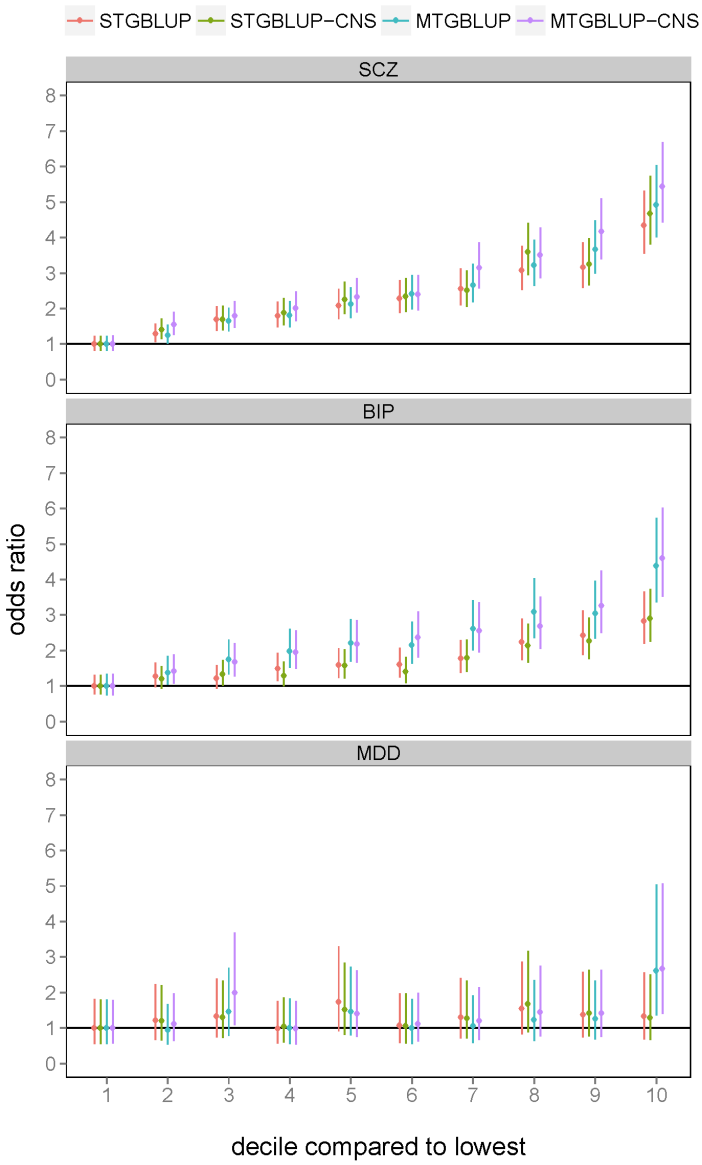

Figure 24: Odds Ratios of Individuals Stratified into Deciles Based on GBLUP Genetic Risk in Independent Samples, using the Decile with the Lowest Risk as the Baseline

The vertical error bars denote 95% CI. We note that the estimates for the different methods are highly correlated, and therefore the vertical error bars cannot be used to infer significance of difference between the methods (see Appendix C).

We also quantified the gain in prediction accuracy from MTGBLUP in terms of sample size. Using recent results on prediction accuracy of polygenic scores derived from quantitative

genetic theory [57,189], we inferred the sample sizes required to achieve the accuracies observed by the methods (**Figure 25**). We assumed prevalence of 1% for schizophrenia, 1% for bipolar disorder and 15% for major depressive disorder. The proportion of cases in the sample was based on the real structure of the discovery data (59% for schizophrenia, 64% for bipolar disorder and 57% for major depressive disorder). The effective number of SNPs was assumed to be 69748 calculated with a weighted SNP method [27]. The observed accuracy was within the theoretical expectation for SCZ and BIP, but not for MDD where the actual predictive power was lower. Accuracy of risk prediction for individual traits benefited from including the correlated disorders. The gain in accuracy of MTGBLUP compared to STGBLUP was equivalent to increasing the sample size for schizophrenia, bipolar disorder and major depressive disorder by ~ 4660 (95% confidence interval: 3110-6270), ~ 5560 (2830-8640) and ~ 10940 (730 – 24440) individuals, respectively (**Figure 25**). Gains in accuracy were even greater using the CNS annotation model (**Table 11**). The 95% confidence interval was obtained according to the sampling error of the difference between the prediction accuracies (Appendix C).

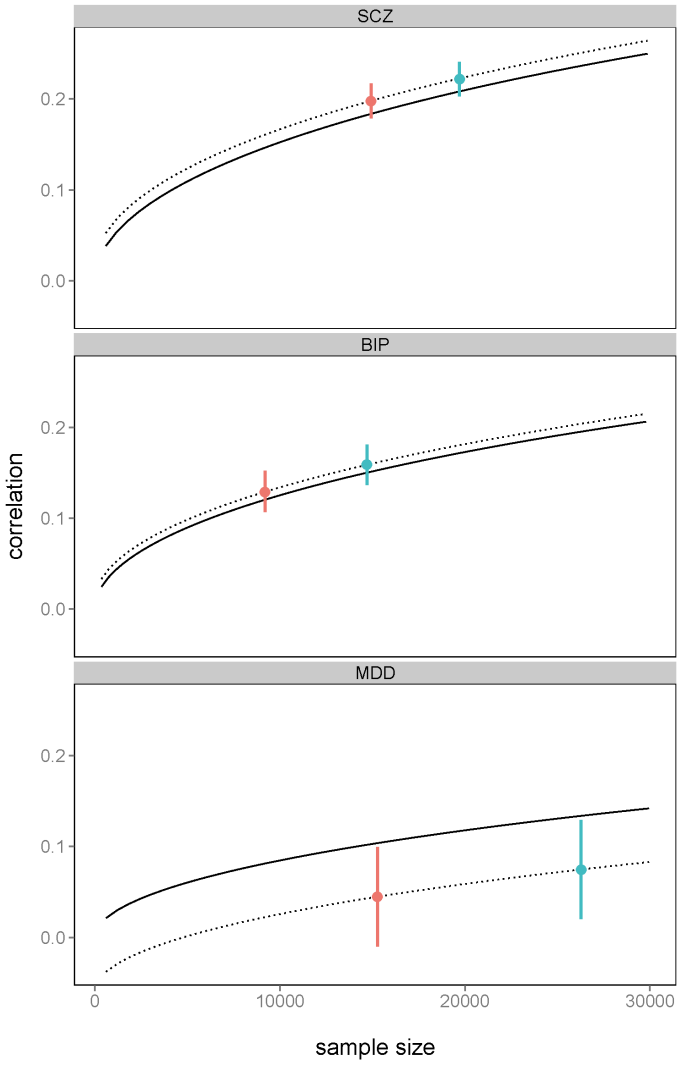

Figure 25: Theoretical and Observed Prediction Accuracy of STGBLUP and MTGBLUP Depending on Sample Size

Theoretical line of prediction accuracy increased with larger sample size (solid line), the observed accuracy achieved by STGBLUP with the actual sample size (red dot), and the observed accuracy achieved by MTGBLUP and inferred sample size (blue dot). The increase from MTGBLUP equates to ~4,660 samples for schizophrenia, ~5,550 samples for bipolar disorder, and ~10,940 for major depressive disorder. The vertical error bars denote 95% CI. We note that the estimates for the different methods are highly correlated, and therefore the vertical error bars cannot be used to infer significance of difference between the methods (see Appendix C).

Table 11: The gain in prediction accuracy from MTGBLUP option in terms of sample size equivalence using STGBLUP

Ninety-five percent confidence interval (CI) is in bracket.

	SCZ	BIP	MDD
MTGBLUP	4660 (3110 – 6270)	5550 (2830 – 8640)	10940 (730 – 24440)
MTGBLUP-CNS	5080 (3520 – 6690)	6220 (3380 – 9380)	11550 (1220 – 25300)

In order to test how sensitive our results on prediction are against population stratification, we re-estimated the prediction accuracy (correlation) removing potential outliers that were ± 6SD, 2 SD, 1.75 SD, 1.5 SD, 1.25 SD or 1 SD away from the mean of the first and second principal component in the validation data set (**Figure 26**, **Figure 27** and **Figure 28**). The accuracy of MTGBLUP and STGBLUP remained stable in all three diseases for which independent datasets were available. Restricting the samples to individuals whose values of the first and second principal component lay within one SD of the mean retained between 51% and 70% of the samples (**Figure 26**, **Figure 27** and **Figure 28**). This shows that the prediction accuracy was not substantially affected by ancestry outliers in the validation dataset.

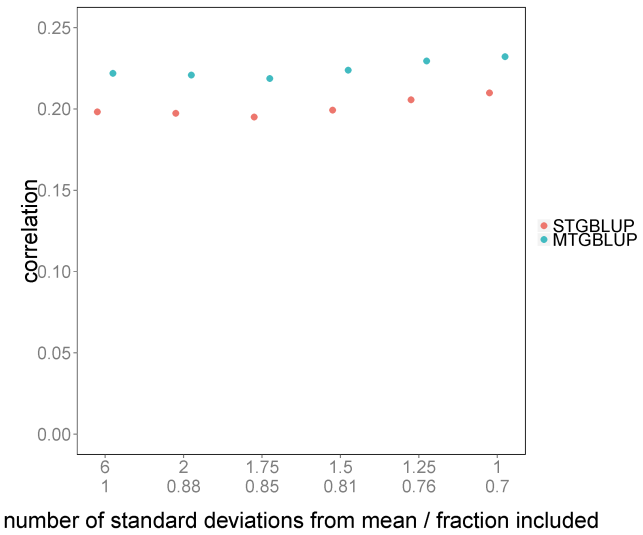

Figure 26: Effect of excluding population outliers – schizophrenia

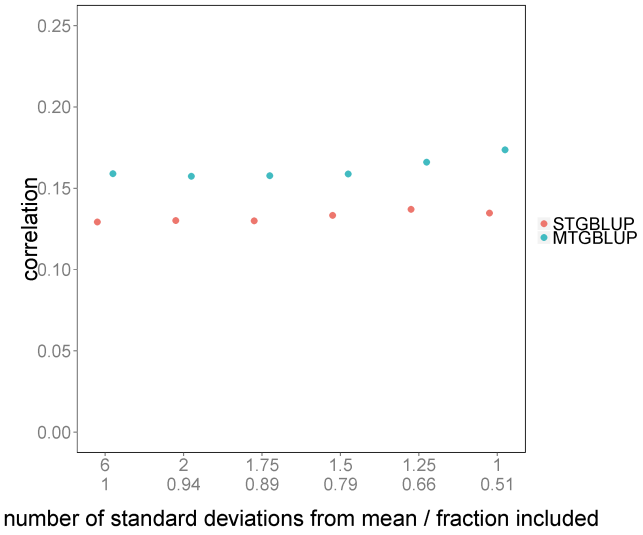

Figure 27: Effect of excluding population outliers – bipolar disorder

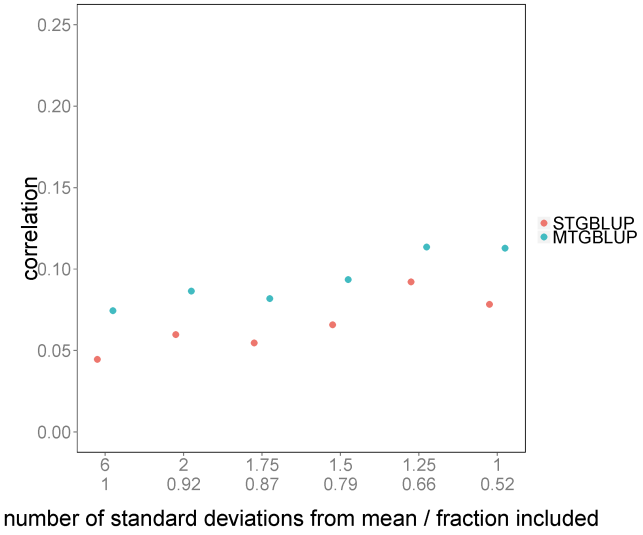

Figure 28: Effect of excluding population outliers – major depressive disorder

*Effect of excluding population outliers on the prediction accuracy from MTGBLUP and STGBLUP. Outliers are defined as points ± 6, 2, 1.75, 1.5, 1.25 and 1 SD from the mean for both the first and second principal components in the independent (**Figure 26**) SCZ, (**Figure 27**) BIP, (**Figure 28**) MDD samples.*

We compared the performance of MTGBLUP with that of bivariate GBLUP (a special case of MTGBLUP). The accuracy of MTGBLUP was significantly higher than bivariate GBLUP except for a MDD risk prediction where the accuracy of MTGBLUP and that of the bivariate model involving SCZ and MDD was not significantly different. (**Table 12** and **Table 13**).

Table 12: Prediction accuracy of bivariate GBLUP (BVGBLUP)

model	Dependent variable	correlation	regression
BVGBLUP (SCZ, BIP)	SCZ	0.220	0.822
BVGBLUP (SCZ, BIP)	BIP	0.156	0.705
BVGBLUP (SCZ, MDD)	SCZ	0.201	0.785
BVGBLUP (SCZ, MDD)	MDD	0.071	0.461
BVGBLUP (BIP, MDD)	BIP	0.133	0.682
BVGBLUP (BIP, MDD)	MDD	0.040	0.263

Table 13: P-values from likelihood ratio test for comparisons among BVGBLUP and MTGBLUP

Likelihood ratio LR = -2 [logL(x1) - logL(x1+x2)]

x1	x2	Dependent variable	p-value
BVGBLUP (SCZ, BIP)	MTGBLUP	SCZ	1.9E-03
BVGBLUP (SCZ, BIP)	MTGBLUP	BIP	7.4E-03
BVGBLUP (SCZ, MDD)	MTGBLUP	SCZ	1.9E-23
BVGBLUP (SCZ, MDD)	MTGBLUP	MDD	0.50
BVGBLUP (BIP, MDD)	MTGBLUP	BIP	2.5E-14
BVGBLUP (BIP, MDD)	MTGBLUP	MDD	0.00056

Psychiatry lags behind other fields of medicine in terms of diagnostic tests that could facilitate early diagnosis and accurate classification of disorders. The considerable heritability of psychiatric disorders implies that the genome contains a large amount of information with potential diagnostic utility. However, the highly polygenic nature of psychiatric disorders makes it very hard to exploit this information, mostly because the effect of each individual locus contributing to disease risk can only be estimated with error, and the size of the error depends on factors such allele frequency, effect size and crucially, sample size.

The genetic correlation between several diseases implies that a SNP contributing to risk of one disease will, on average, also be informative of the risk of the correlated diseases. Here, we have developed a multivariate method that can combine data from an arbitrary number of genetically correlated diseases resulting in better estimates of the disease specific SNP effects and thus generating more accurate predictors of individual risk. Our results demonstrate a significant advantage of incorporating data from multiple correlated diseases compared to single-trait analyses. Our estimates of pairwise genetic correlations obtained in independent datasets reconfirm previous results regarding the extent of genetic correlations between the five psychiatric disorders [10]. External validation demonstrated that the predictive models generalise to other populations, confirming that the correlations reflect pleiotropy between the disorders rather than artefacts.

We used a multiple random effects model that fitted two components, one is due to annotated SNPs and the other is due to the rest of SNPs. The prediction accuracy significantly increased when using an appropriate gene set. For example, the gain in predictive accuracy in terms of sample size equivalence increased from 4660 to 5080 for schizophrenia, from 5550 to 6220 for bipolar disorder, and from 10940 to 11550 for major depressive disorder when using the CNS genes annotation [10,180] (**Table 11**). This demonstrates that the multiple random effects model in MTGBLUP can be useful especially for psychiatric disorders where prediction accuracy is hardly improved by other advanced methods [163,164].

Zhou and Stephens [28] recently introduced a multivariate linear mixed model algorithm that is particularly suited for genome-wide association studies. Their method requires that

multiple traits are measured on the same individual or that the level of missingness is sufficiently small so that missing phenotypes can be imputed. However, this algorithm is not useful when phenotypes are collected from independent data sets as in the PGC data where dependent variables are totally missing for the other four traits as is typical of disease ascertained cohorts. Moreover, the efficiency of Zhou and Stephens' algorithm substantially decreases when fitting multiple random effects (e.g. the annotation model).

Korte et al. (2012) [190] proposed a similar model to MTGREML using ASReml [191] that is as flexible as our method in that it can handle partial overlapping or disjoint sets of phenotypes. However, our algorithm is different from that used in ASReml and is much more efficient when using genomic data [175] (see Appendix A). Moreover, Korte et al. did not explore their method with respect to improvements in risk prediction.

Even though sensitivity and specificity of genetic diagnostics to predict an individual's risk of psychiatric disorders are generally low, genetic risk scores can still be a valuable tool for research to stratify a heterogeneous population in groups with shared 'genomic' characteristics. It was suggested that psychiatric diagnoses encompass several clinically similar phenotypes with distinct pathophysiology and that stratification according to individual heterogeneity is an important requirement for the development of treatments targeted at specific disease subtypes [111,156,174]. Our proposed multivariate approach with the annotation model is a flexible and powerful tool for such stratification. The MTGREML and MTGBLUP package and documentation are publicly available online (see Web Recourses), which we anticipate will be implemented into the GCTA package [26]. Using a CPU running at 2.2 GHz, analysing 58128 sample with 5 disjoint set of phenotypes (e.g. the PGC data) takes ~ 7 hours per each iteration in MTGREML. Convergence is usually achieved within 10 iterations. The virtual memory required for such data is ~ 45 GB. Good starting values (probably from single trait GREML [190]) can reduce the number of iterations to convergence and our software has the option to provide starting values. The computational time increases cubically with sample size, e.g. analysing sample size of 10000 takes a few minutes per each iteration. Our software provides a parallelisation option that can reduce computational burden substantially, for example speed is increased by a factor of ten when using 20 CPUs. The number of traits hardly affects running time if phenotypes are non-overlapping.

Appendix A

Average of Hessian and Fisher information matrix for the multivariate model

The log likelihood of the multivariate model is,

$$\ln L = \frac{1}{2}[\ln|\mathbf{V}| + \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y}]$$

where ln is the natural log, and | | the determinant of the associated matrices. The projection matrix is defined as $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ with $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{X}_n \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}$.

The Newton-Raphson algorithm obtains the MTGREML estimates using the following equation [192].

$$\Theta^{(k+1)} = \Theta^{(k)} + \left(\mathbf{H}^{(k)}\right)^{-1} \frac{\partial L}{\partial \Theta}|\Theta^{(k)} \qquad \text{(A1)}$$

where Θ is a column vector of estimated variance components, *k* is the iteration round, $\frac{\partial L}{\partial \Theta}$ is a column vector of the first derivatives of the log likelihood function with respect to each variance component, and **H** is the Hessian matrix which consists of the second derivatives of the log likelihood function with respect to the variance components. In Fisher's scoring method, the inverse of the Hessian matrix in (A1) is replaced by its expected value [192].

$$\Theta^{(k+1)} = \Theta^{(k)} + \left(F^{(k)}\right)^{-1} \frac{\partial L}{\partial \Theta}|\Theta^{(k)} \qquad \text{(A2)}$$

The derivation of the Hessian matrix and the Fisher information matrix has been described in several studies [192,193]. The Hessian matrix for the multivariate model is

$$\mathbf{H} = \frac{\partial^2 L}{\partial \sigma_i^2 \partial \sigma_j^2} = \frac{1}{2}\left[tr\left(\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\right) - \mathbf{y}'\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\mathbf{P}\mathbf{y}\right] \qquad \text{(A3)}$$

where **y**, **P** and **V** are defined in the section 'Multivariate linear mixed model' in the main text. The Fisher information (F) matrix is

$$\mathbf{F} = E\left(\frac{\partial^2 L}{\partial \sigma_i^2 \partial \sigma_j^2}\right) = \frac{1}{2}\left[tr\left(\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\right)\right] \qquad \text{(A4)}$$

Gilmour et al. (1995) [191] and Johnson and Thompson (1995) [194] used the average of the **H** and **F** that was estimated based on Henderson's mixed model equation (MME) [195]. The MME-based average information algorithm is efficient particularly when covariance

structure fitted in the model is sparse. Lee and van der Werf (2006) [175] introduced the direct average information algorithm where average information matrix was derived directly from the **V** and **P** matrix. When using non-zero elements of covariance structure, this direct average information algorithm is much more efficient than the MME-based average information algorithm. The equation for the iterative AI algorithm is

$$\Theta^{(k+1)} = \Theta^{(k)} + \left(\mathbf{AI}^{(k)}\right)^{-1} \frac{\partial L}{\partial \Theta} | \Theta^{(k)}$$

where **AI** is the average information matrix and that for multivariate model can be written as

$$\mathbf{AI} = \frac{1}{2}[\mathbf{y}' \frac{\partial \mathbf{V}}{\partial \sigma_i^2} \mathbf{P} \frac{\partial \mathbf{V}}{\partial \sigma_i^2} \mathbf{P}\mathbf{P}\mathbf{y}]$$

The first derivative for each variance covariance component *i* can be obtained as [192,193]

$$\frac{\partial L}{\partial \sigma_i^2} = -\frac{1}{2} tr(\mathbf{P} \frac{\partial \mathbf{V}}{\partial \sigma_i^2}) + \frac{1}{2} \mathbf{P} \frac{\partial \mathbf{V}}{\partial \sigma_i^2} \mathbf{P}\mathbf{y}$$

Appendix B

Reaction norm model to test heterogeneity across populations classified by the ancestry principal component

Reaction norm models have been used in ecology and evolution to study genotype x environment interaction [187,188]. Genotype x environment interaction (G x E) means that different genotypes respond different to environmental changes, i.e. norms of reaction. In the model, a random intercept and a random slope, as covariance functions, are estimated that can describe genetic and phenotypic variation across different environments. The slope of the reaction norm is often called phenotypic plasticity or environmental sensitivity. The amount of variation in slope in the population indicates the extent of G x E [187,188]. Here, we describe a reaction norm model to test heterogeneity across populations. We group each sample set into four populations by splitting them into the four quartiles of the first ancestry principal component. Whereas typically reaction norm models would compare samples with different categories of environmental factors to each other, we use the model to compare the samples in different principal component quartiles to each other. We limit our interpretation to heterogeneity across the groups, and do not speculate about potential causes like G x E or G x G interaction. We apply the model to each disorder of the PGC data. Incorporating population difference among sample, the linear mixed model can be rewritten as

$$y_{ij} = b_{ij} + g_{ij} + e_{ij}$$

where y_{ij} is the observation for individual i in population class j (j=1, …, P where P is the number of populations classified by the ancestry principal component, in our case 4), b_{ij} is fixed effects, g_{ij} is genetic effects and e_{ij} is residual effects. We applied a reaction norm model to fit functions of the ancestry principal component as covariables using Legendre polynomials.

$$\mathrm{y_{ij}} = \mathrm{b_{ij}} + \sum_{m=0}^{k-1} \alpha_{im} f_m(p_{ij}) + e_{ij}$$

p_{ij} is the average of the ancestry principal components in the j th population class containing individual i, $f_m(p_{ij})$ is the m th Legendre polynomials evaluated for p_{ij}, α_{im} is the m th genetic random regression coefficients for the i th individual, and k is the orders of fit. The genetic covariance between individual i in population class j and i' in population class j' is

$cov(\mathrm{g_{ij}}, \mathrm{g_{i'j'}}) = \sum_{m=0}^{k-1} \sum_{l=0}^{k-1} f_m(p_{ij}) f_l(\mathrm{p_{i'j'}}) cov(\alpha_{im}, \alpha_{il})$.

This can be written in a matrix from as

$\mathbf{V}_g = \mathbf{FKF}'$

where **F** is the matrix of Legendre polynomials evaluated at given ancestry principal components and **K** is the covariance coefficient matrix consisting of random regression coefficients, i.e.

$$\mathbf{K} = cov(\alpha_{im}, \alpha_{il}) = \begin{bmatrix} var(\alpha_0) & \cdots & cov(\alpha_0, \alpha_k) \\ \vdots & \ddots & \vdots \\ cov(\alpha_k, \alpha_0) & \cdots & var(\alpha_k) \end{bmatrix}$$

The optimal order of the polynomial was determined with a likelihood ratio test by comparing the likelihood of models with higher order to the null model with k=1.

Appendix C

Estimating the sampling error of the difference between prediction accuracies (correlations)

It is assumed that there are three normalised variables with the covariance structure as below, mimicking the MTGBLUP, STGBLUP and outcome variable.

Table 14: Multi-trait – single-trait correlations

	m	s	y
m	1	0.927	0.222
s	0.927	1	0.189
y	0.222	0.198	1

We are interested in estimating the sampling error of the difference between cor(m,y) and cor(s,y). The sampling variance of the difference (σ^2_{d}) can be expressed as

$$\sigma^2_{\mathrm{d}} = \sigma^2_{\mathrm{cor(m,y)}} + \sigma^2_{\mathrm{cor(s,y)}} - 2r\sigma_{\mathrm{cor(m,y)}}\sigma_{\mathrm{cor(s,y)}} \quad \text{(C1)}$$

where $\sigma^2_{\mathrm{cor(m,y)}}$ is the sampling variance of cor(m, y) and $\sigma^2_{\mathrm{cor(s,y)}}$ is the sampling variance of cor(s, y) and r is the correlation between cor(m, y) and cor(s, y). We show here that r is approximately equal to cor(m, s).

With N records for each variable, correlations among the variables can be written as

$$\mathrm{cor(m,y)} = E(m\,y) = \frac{1}{N}\sum_{i=1}^{N} m_i y_i$$

$$\mathrm{cor(s,y)} = E(s\,y) = \frac{1}{N}\sum_{i=1}^{N} s_i y_i$$

$$\mathrm{cor(m,s)} = E(m\,s) = \frac{1}{N}\sum_{i=1}^{N} m_i s_i$$

For T replicates, the expected value of the product of cor(m,y) and cor(s,y) can be written as

$E[E(m\ y)E(s\ y)] = \frac{1}{T}\sum_j^T \left[\left(\frac{1}{N}\sum_{i=1}^N m_i y_i\right)\left(\frac{1}{N}\sum_{i=1}^N s_i y_i\right)\right]_j =$

$\frac{1}{T}\sum_j^T \left[\left(\frac{m_1 y_1}{N} +, \ldots, + \frac{m_N y_N}{N}\right)\left(\frac{s_1 y_1}{N} +, \ldots, + \frac{s_N y_N}{N}\right)\right]_j$.

If m and s are uncorrelated, this reduces to

$E[E(m\ y)E(s\ y)] = E[E(\mathrm{m\ y})]\ E[E(s\ y)]$

If m and s are correlated, there is an additional term,

$E[E(m\ y)E(s\ y)] \cong \frac{1}{T}\sum_j^T \left[\left(\frac{1}{N^2}\sum_{i=1}^N m_i s_i\right)\right]_j + E[E(\mathrm{m\ y})]\ E[E(s\ y)] = \frac{1}{T}\sum_j^T \left[\frac{1}{N}E(\mathrm{m\ y})\right]_j +$

$E[E(\mathrm{m\ y})]\ E[E(s\ y)] = \frac{1}{N}E[E(m\ s)] + E[E(m\ y)]E[E(s\ y)]$.

Therefore,

$cov[E(m\ y), E(s\ y)] = \ E[E(m\ y)E(s\ y)] - \ E[E(m\ y)]\ E[E(s\ y)] \cong \frac{1}{N}E[E(m\ s)]$

With $var[E(m\ y)] \cong var[E(s\ y)] \cong \frac{1}{N}$, the correlation between cor(m, y) and cor(s, y) (r) can be approximated as

$cor[E(m\ y), E(s\ y)\ ] \cong E[E(m\ s)] \cong cor(m, s)$.

This expression was checked and validated by simulations (result not shown).

Here we have shown that Equation (C1) can be used to estimate the sampling variance (the square of the standard error) of the difference in correlation between the STGBLUP and MTGBLUP predictors (which are themselves correlated with each other) and the outcome variable (the adjusted phenotype). This allows us to estimate the 95% confidence interval of the increase in correlation which MTGBLUP achieves over STGBLUP. Note that since the two predictors are correlated, this is a smaller confidence interval than that of the correlation between MTGBLUP and the outcome variable (which is shown in **Figure 24** and **Figure 25**). Using the method described above, we can transform the confidence interval from the correlation scale to the sample size scale, to get estimates of the effective increase in sample size achieved by MTGBLUP (**Table 11**).

Appendix D

It has been pointed out that the reaction norm model, which in this work has been used to detect genetic heterogeneity between discovery and validation set along the first principal component, bears resemblance to the genotype clustering approach of Chapter 3. This section explores what the similarities and differences between these two approaches are.

Estimating genetic correlations across different populations can be tricky. Even when the true effect sizes for the trait are the same across populations, differences in LD structure can lead to estimates lower than one. Furthermore, identical effect sizes across populations don't necessarily imply that each SNP explains the same amount of variance in the two populations, since allele frequencies can differ. This has motivated the distinction between transethnic genetic correlation and transethnic genetic impact correlation [38]. In this Chapter, we showed that when we group individuals in our discovery and validation set into four groups each, according to their first principal component, which captures the direction of the largest ancestry variation in the (European) data, the genetic effects differ for more distant groups, but not for groups which fall into the same PC 1 region. This can be caused either by differences in the true effect sizes among more distant groups or by differences in the LD structure.

In the analyses of Chapter 3, the goal was not to start with two groups of individuals (discovery and validation) and to test whether there is genetic heterogeneity between the two groups, conditional on ancestry. Rather, we assumed that there is genetic heterogeneity with respect to a phenotype and wanted to test whether clustering will detect this heterogeneity or whether it will pick up other effects, such as ancestry or population stratification. What the k-means clustering algorithm picks up depends very much on the first principal component of the data used for clustering. Generally, the clusters will form along the direction of the largest variance in the data. If the clustering had been performed on all SNPs in the data, this would simply be the first ancestry principal component, and hence the clustering would only be informative of case status to the extent to which case status is confounded with PC1. To prevent the clustering from picking up ancestry, we selected only SNPs associated with the phenotype (after correction for PCs). Ideally, the first principal component in this subset should not reflect ancestry or stratification, but rather case status, which would lead to clusters that can distinguish both types of cases.

To summarise, in both the Chapter 3 and Chapter 4 analysis, there are groups of individuals with a certain genetic correlation for a given phenotype. In Chapter 4 the first principal component was fixed and was used to define groups and for which the genetic correlation

was then determined. In Chapter 3 the goal was to select a subset of SNP such that the first principal component would recover the two groups with a genetic correlation of less than one.

Acknowledgments

This study was supported by the Australian Research Council (FT0991360 and DE130100614) and the National Health and Medical Research Council (613608, 1011506 and 1047956). The Psychiatric Genomics Consortium is supported by National Institute of Mental Health (NIMH) grant U01 MH085520. We acknowledge the funding that supported the Swedish schizophrenia study (NIMH R01 MH077139), the Stanley Center for Psychiatric Research, the Sylvan Herman Foundation, the Karolinska Institutet, Karolinska University Hospital, the Swedish Research Council, the Stockholm County Council, the Söderström Königska Foundation, and the Netherlands Scientific Organization (NWO 645-000-003). Statistical analyses were carried out on the Genetic Cluster Computer (see Web Resources), which is financially supported by the Netherlands Scientific Organization (NOW; 480-05-003). The GenRED GWAS project was supported by NIMH R01 Grants MH061686 (DF Levinson), MH059542 (W Coryell), MH075131 (WB Lawson), MH059552 (JB Potash), MH059541 (WA Scheftner) and MH060912 (MM Weissman).

Web Recourses

Psychiatric Genomics Consortium
https://pgc.unc.edu/

Genetic Cluster Computer
http://www.geneticcluster.org/

Genome-wide Complex Trait Analysis (GCTA),
http://www.complextraitgenomics.com/software/gcta

Multi-trait GREML and GBLUP analysis (MTG),
http://www.cnsgenomics.com/software/
https://github.com/uqrmaie1/mtgblup

5

Chapter 5: Improving genetic prediction by leveraging genetic correlations among human diseases and traits

Chapter submitted

Chapter 5: Improving genetic prediction by leveraging genetic correlations among human diseases and traits

Robert M Maier[1,2], Zhihong Zhu[2], Sang Hong Lee[1,3], Maciej Trzaskowski[2], Douglas M Ruderfer[4], Eli A Stahl[5], Stephan Ripke[6,7,8], Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Naomi R Wray[1,2], Jian Yang[1,2], Peter M Visscher[1,2] & Matthew R Robinson[2,9]

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.
[2]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia.
[3]School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia
[4]Division of Genetic Medicine, Department of Medicine, Psychiatry and Biomedical Informatics, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
[5]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]Stanley Center for Psychiatric Research, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA
[7]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[8]Department of Psychiatry and Psychotherapy, Charité, Campus Mitte, Berlin, Germany
[9]Department of Computational Biology, University of Lausanne, 1011 Lausanne, Switzerland

Abstract

Genomic prediction has the potential to contribute to precision medicine where diagnosis and treatment are tailored to individuals on the basis of their genetic risk of disease. However, current genetic predictors of complex human disorders and quantitative traits are generally characterized by low prediction accuracy, which limits their utility. Here, theory and simulation study are used to demonstrate that widespread pleiotropy among phenotypes can be utilized to improve genomic risk prediction. We show how a genetic predictor can be created as a weighted index that combines published genome-wide association study (GWAS) summary statistics across many different traits. We apply this framework to predict risk of schizophrenia and bipolar disorder in the Psychiatric Genomics consortium data, finding substantial heterogeneity in prediction accuracy increases across cohorts. For six additional phenotypes in the UK Biobank data, we find increases in prediction accuracy ranging from 0.7% for height to 47% for type 2 diabetes, when using a multi-trait predictor that combines published summary statistics from multiple traits, as compared to a predictor based only on one trait.

Introduction

Personalized medicine, in which genetic testing is the basis for informing future health status and determining intervention, is effectively applied for a number of monogenic disorders [196]. For common complex disorders, which are those that are underlain by multiple genetic and environmental factors [7], predictive genetic testing that can discriminate individuals who are most at risk is currently limited, mainly because much of the genetic variation remains poorly understood [51,56]. Improving the accuracy of genetic risk prediction has the potential to (i) prospectively identify individuals at increased risk of disease, thus informing early interventions, and (ii) aid diagnosis for diseases where current diagnostic approaches are imperfect [197]. While genome-wide association studies (GWAS) of increased sample size will continue to unravel the role of genetic factors for complex diseases [11], improved prediction models are also required to maximize the accuracy of a risk predictor.

GWAS use linear regression to independently estimate the effects of single nucleotide polymorphisms (SNPs) across the genome, and commonly, these estimated SNP effects

are then used to create a genetic risk predictor in independent samples [24,54,198]. However, this approach is not optimal because it either ignores LD between markers, or accounts for LD by discarding potentially informative SNPs [70]. Prediction accuracy of complex phenotypes can be improved by methods that jointly estimate the SNP associations to obtain best linear unbiased predictors (BLUP) of the SNP effects within a linear mixed model (LMM) approach [24,199,200]. A multi-trait extension of the LMM approach, yielding multivariate BLUP (MT-BLUP) predictors of the SNP effects, can further improve prediction accuracy when phenotypes are genetically correlated, because measurements on each trait provide information on the genetic values of the other correlated traits [66,170,201,202]. MT-BLUP has been shown to improve prediction accuracy for genetically correlated common psychiatric disorders when combining individual-level data across independent data sets [66,168]. However, the application of MT-BLUP to complex common disorders is limited as combining individual-level genotype-phenotype data across case-control studies of all complex diseases is generally not feasible due to data protection concerns and restrictions on data sharing.

Here, we overcome this limitation by developing a framework that combines publically available GWAS summary statistics across multiple studies of different traits together in a weighted index to generate approximate multi-trait summary statistic BLUP (wMT-SBLUP) predictors (**Table 15**). We show through theory and simulation study that MT-BLUP predictors, which traditionally require individual-level phenotype-genotype data for all traits, can be approximated accurately by wMT-SBLUP predictors in a computationally efficient manner using only summary statistic data and an independent genomic reference sample. We apply this approach to multiple phenotypes in the Psychiatric Genomics Consortium (PGC) and the UK Biobank data and show increased prediction accuracy as compared to a single trait predictor.

Table 15: SNP-heritability estimates and sample size for each summary statistics trait, as well as matched UK Biobank traits

Trait	h^2	SE	Median N	PMID	UKB matched	UKB N
ADHD	0.17	0.1	5422	20732625		
Agreeableness	0.02	0.02	17375	21173776		
Alzheimers	0.05	0.03	54162	24162737		
Autism	0.41	0.05	10610	23453885		
Bipolar	0.27	0.02	29031	24280982	mania/bipolar disorder/manic depression	301
Birth Length	0.17	0.02	22145	25281659		
Birth Weight	0.11	0.02	26836	23202124		
BMI	0.14	0.01	233723	25673413	bmi	112027
CAD	0.07	0.01	184305	21378990	angina	3847
Childhood Obesity	0.42	0.05	13848	22484627		
Conscientiousness	0.08	0.03	17375	21173776		
Crohn's disease	0.52	0.06	20883	26192919		
Depression	0.18	0.03	16610	23453885	depression	4698
Diabetes	0.09	0.01	63390	22885922	diabetes	4978
Education Years	0.11	0.01	216772	23722424		
Extraversion	0.03	0.03	17375	21173776		
Glucose	0.15	0.03	38422	22581228		
Head Circumference	0.21	0.05	10761	22504419		
Height	0.35	0.02	252190	20881960	height	112147
Inflammatory Bowel Disease	0.35	0.04	34652	26192919		
Insulin	0.1	0.02	33823	22581228		
IQ	0.19	0.03	17989	23358156	fluid_intelligence_score	36093
MND	0.04	0.02	36052	27455348		
Neuroticism	0.04	0.01	63661	25993607		

Openness	0.11	0.03	17375	21173776		
Osteoporosis Femur	0.13	0.02	49988	26367794		
Pubertal Growth	0.44	0.05	13955	23449627		
Rheumatoid Arthritis	0.23	0.06	25500	20453842		
Schizophrenia	0.25	0.01	150064	25056061	schizophrenia	131
Smoking	0.08	0.01	74035	20418890		
Tanner	0.12	0.05	9915	24770850		
Triglycerides	0.27	0.06	90981	20686565		
Ulcerative Colitis	0.27	0.04	27432	26192919		
Waist Hip Ratio	0.09	0.01	142471	25673412		

Results

Overview of the approach. Standard GWAS summary statistics are OLS estimates of the SNP effects and do not have optimal properties for prediction [199]. Even when LMM association analysis is used, the estimated SNP effects still represent marginal effects, and not effects conditional on other SNP, which is what is desirable for prediction [31]. Previous studies have shown how OLS summary statistics can be reanalysed in a mixed model framework to produce approximate BLUP predictors (summary statistic BLUP: SBLUP) [67,68,203]. We first extend this approach to a multi-trait framework (MT-SBLUP) and find a computational limitation associated with the inversion of a SNP-by-SNP-by-trait matrix. To overcome this, we then derive theory to show how single trait predictors with BLUP properties can be combined together in a weighted index to generate predictors with equivalent properties to those gained from a MT-BLUP analysis (Methods, **Figure 29**, **Table 16**).

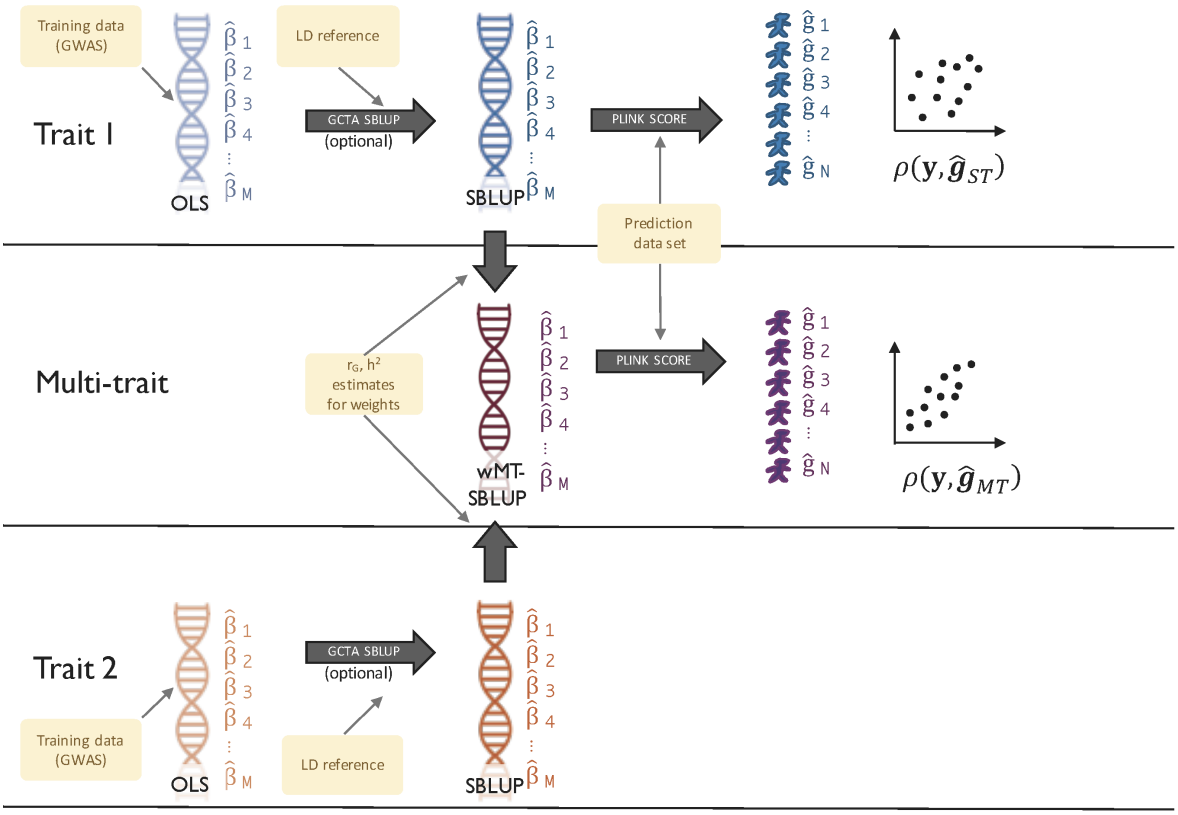

Figure 29: Data and programs used to create predictors

GCTA and an external reference data set is used to turn GWAS effect estimates into SBLUP effect estimates; two or more traits are combined to create wMT-SBLUP estimates; these are then converted into individual predictors using PLINK.

Table 16: Terminology to refer to different types of predictors

OLS: Ordinary Least Squares. The most common GWAS methodology to estimate SNP effects is to estimate the effect sizes of one SNP at a time. BLUP: Best Linear Unbiased prediction. SNP effects are estimated simultaneously for all SNPs. The estimates depend on the other SNPs included in the analysis, since the contribution from correlated SNPs with be shared between them.

		Single-trait starting point		
		Full Genotype-Phenotype data	**GWAS OLS summary statistics**	**GWAS OLS summary statistics converted to BLUP summary statistics**
Multi-trait conversion	**None (Single Trait)**	BLUP	OLS	SBLUP
	Full Multiple-Trait	MT-BLUP	-	MT-SBLUP
	Approximate Multiple Trait from weighting of single trait predictors	wMT-BLUP	wMT-OLS	wMT-SBLUP

Consider two genetically correlated traits for which we have individual-level genetic predictors with BLUP properties. For each individual, *i*, and focal trait of interest, *f*, we have a genetic prediction ($\hat{\mathbf{g}}_{\mathrm{BLUP}_{i,k}}$) for each trait, *k*, that we can combine together using the index weights, $\mathrm{w}_{i,k}$, for each $\hat{\mathbf{g}}_{\mathrm{BLUP}_{i,k}}$effect to produce a weighted multi-trait BLUP genetic predictor:

$$\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}_{i,f}} = \sum_k \mathrm{w}_{i,k}\hat{\mathrm{g}}_{\mathrm{BLUP}_{i,k}} = \mathbf{w}_i'\hat{\mathbf{g}}_{\mathrm{SBLUP}_i} \quad [1]$$

In the methods section we show that the optimal index weights can be calculated as:

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} R_1^2 & \frac{r_G R_1^2 R_2^2}{\sqrt{h_1^2 h_2^2}} \\ \frac{r_G R_1^2 R_2^2}{\sqrt{h_1^2 h_2^2}} & R_2^2 \end{bmatrix}^{-1} \begin{bmatrix} R_1^2 \\ r_G \sqrt{\frac{h_1^2}{h_2^2}} R_2^2 \end{bmatrix} \quad [2]$$

where h_k^2 is the SNP-heritability of trait *k* (proportion of phenotypic variance explained by genome-wide SNPs), r_G is the genetic correlation between trait *k* and the focal trait, andR_k^2 is the expected squared correlation between a phenotype and a BLUP predictor, calculated as:

$$R_k^2 = \frac{h_k^2}{1+M_{eff}\frac{1-R_k^2}{N_k h_k^2}} \quad [3]$$

where M_{eff} is the effective number of chromosome segments and N_k is the sample size of trait *k*. These weights will ensure that the contribution of each added trait is approximately proportional to the square root of its sample size, its SNP-heritability and its genetic correlation with the focal trait (trait 1), while accounting for different variances of single trait BLUP predictors.

Both h_k^2 and r_G can be estimated from GWAS summary statistics [9,32]. Following [67], individual-level genetic predictors with BLUP properties can also be obtained from GWAS summary statistics ($\hat{\mathbf{g}}_{\mathrm{SBLUP}_k}$, where SBLUP represents summary statistic approximate BLUP). Therefore for any given trait, genetic predictors with BLUP properties ($\hat{\mathbf{g}}_{\mathrm{SBLUP}_k}$) can be created from GWAS summary statistics and these can then be placed in a weighted index to produce approximate multi-trait summary statistic BLUP (wMT-SBLUP) predictors, using only LD score regression and an independent reference sample. This approach approximates MT-BLUP predictors without the need for individual-level phenotype-genotype data for all traits, enabling prediction accuracy to be improved by fully utilizing all of the publically available GWAS summary statistic data.

Simulation study. We first conducted a simulation study using observed SNP genotype data to confirm the expectations from our theory. We show through theory (Online methods) that a wMT-SBLUP genetic predictor has the same expected prediction accuracy as one created from a multivariate mixed effects model (multi-trait BLUP: MT-BLUP) if the linkage disequilibrium among SNP markers in the individual-level analysis is well approximated by a reference genotype panel. (Online Methods). We demonstrate that a wMT-SBLUP predictor increases prediction accuracy over a single-trait predictor, with the magnitude of increase being proportional to the ratio of the SNP-heritability of the added traits relative to that of the predicted trait, the sample size of the added traits relative to that of the predicted trait, and the genetic correlation between the added traits and the predicted trait (**Figure 30**, **Figure 31** and **Figure 32**).

We also provide a theoretical expectation for the loss in prediction accuracy that occurs when using an independent reference sample to compute SBLUP effects compared to a

predictor based on BLUP effects (see Methods), and we detail the loss of prediction accuracy in our simulation study (**Figure 30**, **Figure 31** and **Figure 33**).

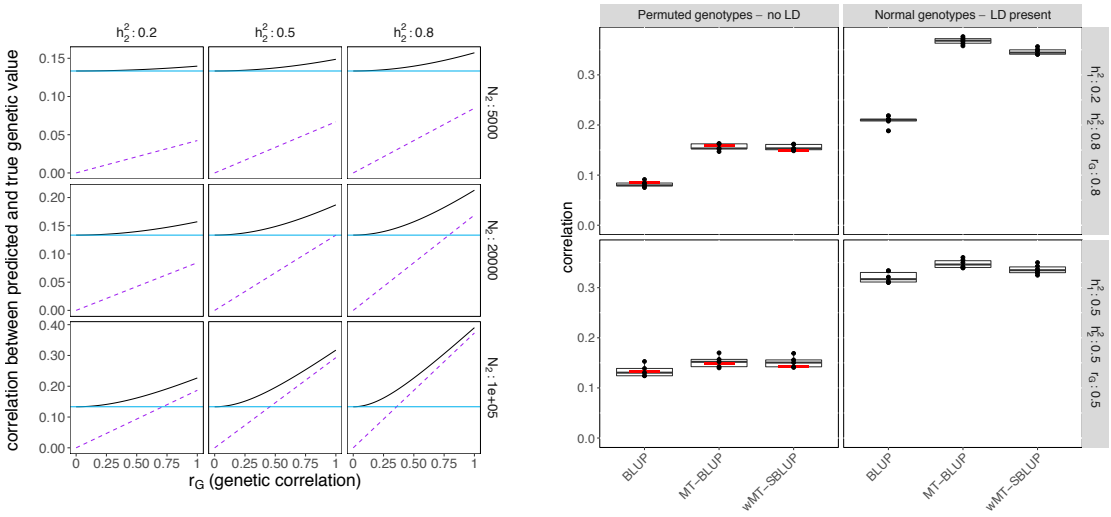

Figure 30: Improving prediction accuracy using information from multiple traits

(a) Expected gain from multi-trait vs cross-trait predictors as a function of r_G. Two traits are considered. The first trait has a sample size of 20,000 and a SNP-heritability of 0.5. The sample size and SNP-heritability of the second trait vary between panels. The blue line shows the expected prediction accuracy of a single-trait predictor. The black line shows the expected prediction accuracy of a multi-trait predictor. The purple line shows the expected prediction accuracy of a cross-trait predictor (using only trait 2 to predict trait 1). The advantage of a multi-trait predictor over a cross-trait predictor decreases with increasing r_G, h^2 and sample size of the second trait.

(b) Simulations results. Prediction accuracy is shown as correlation between simulated genetic value and predicted phenotype of individuals. Genotypes from European individuals in the GERA cohort were used for simulation. Boxplots show results across 6 replicates. In the left panels the LD structure was removed by permuting dosage values for each SNP across all individuals. In the right panels the original genotypes were used for simulation. Expected prediction accuracies were derived for the case of unlinked genotypes and are shown as red horizontal bars. In each section the prediction accuracy of three predictors is shown: (1) single trait BLUP, (2) multi-trait BLUP (MT-BLUP), (3) weighted approximate BLUP (summary statistic based multi-trait predictor: wMT-SBLUP). Simulation in genotypes without LD results in prediction accuracies which conform to expectations. In the presence of LD individual level correlations are higher because the effective number of markers is smaller.

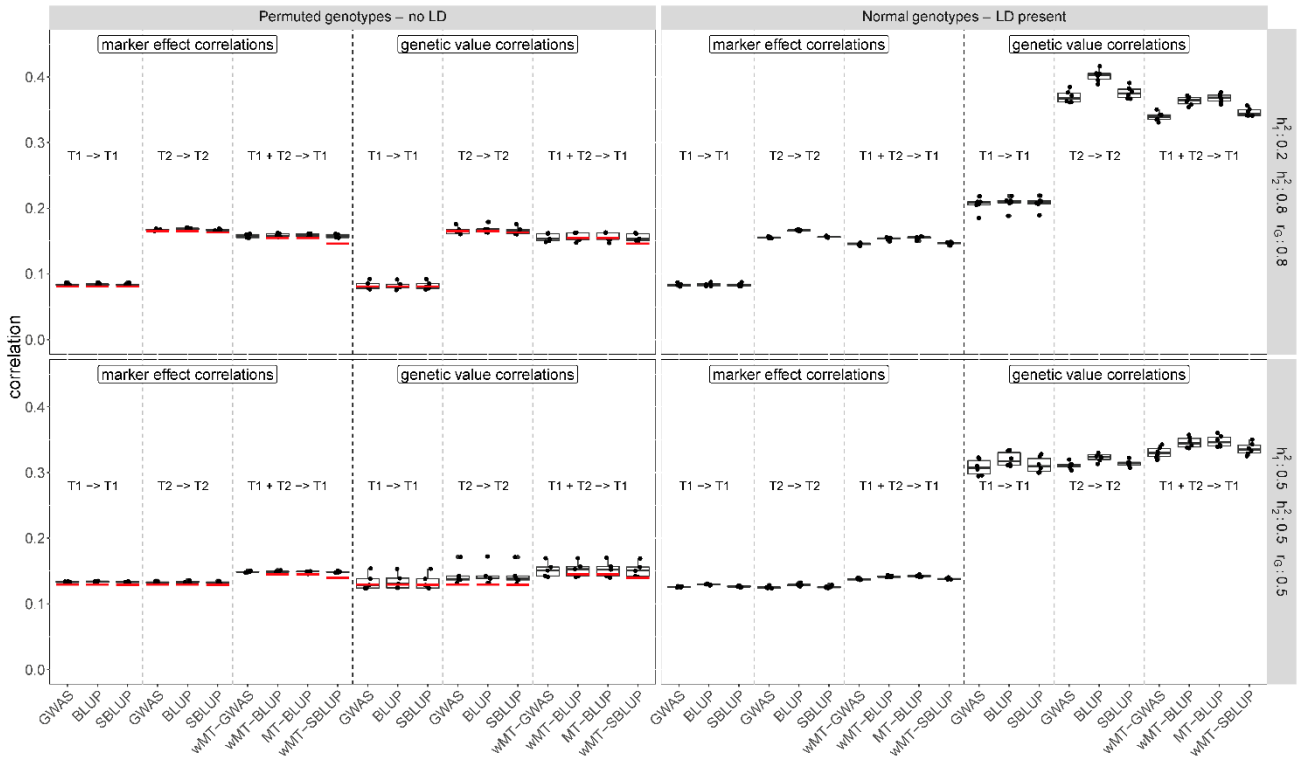

Figure 31: Extended simulation results

Each of the four panels has 10 predictors. For all predictors the correlation of simulated and estimated effect sizes are shown in the left half and the correlation between simulated genetic value and genetic predictor are shown in the right half. Predictors are ordered as follows: 1 to 3, predicting trait 1, while training on trait 1 (1: GWAS, 2: BLUP, 3: SBLUP); 4 to 6, predicting trait 2, while training on trait 2 (4: GWAS, 5: BLUP, 6: SBLUP); 7 to 10, predicting trait 1, while training on trait 1 and 2 (multi trait predictor; 7: wMT-GWAS, 8: wMT-BLUP, 9: MT-BLUP; 10: wMT-SBLUP)

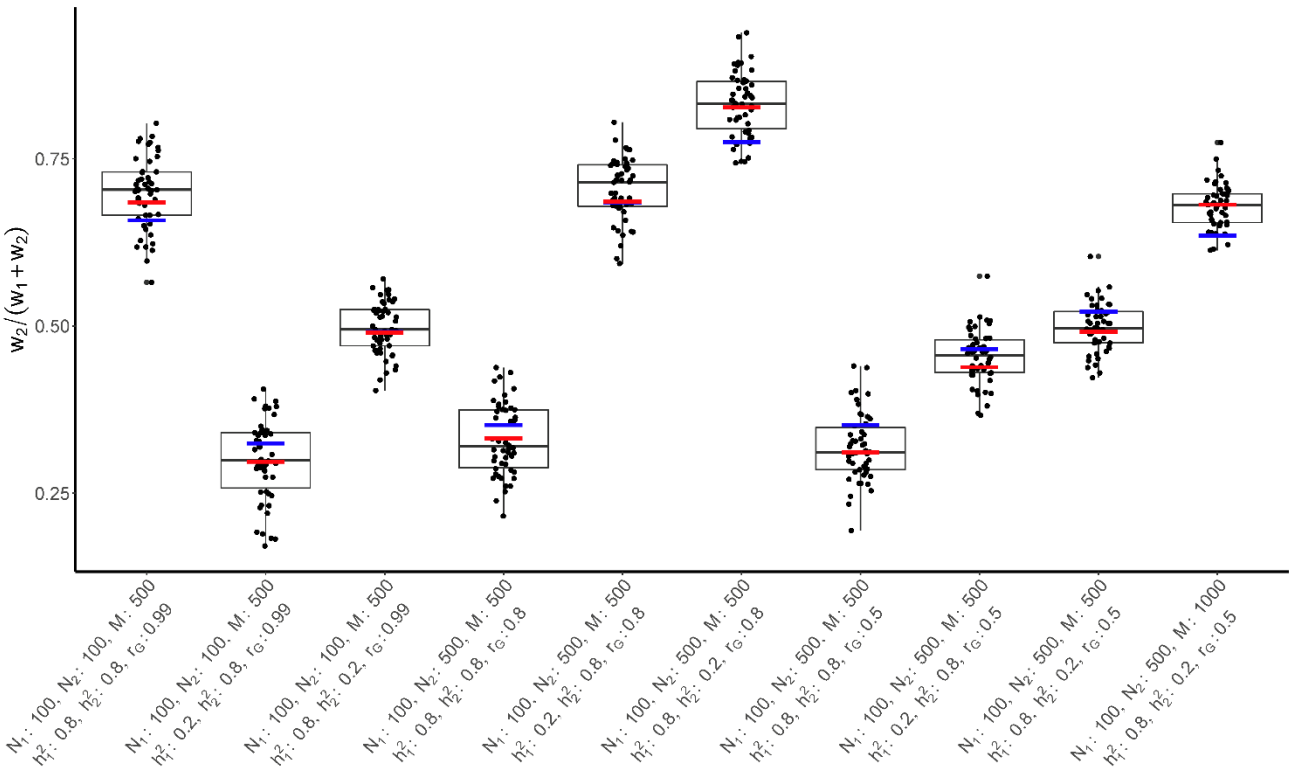

Figure 32: Theoretically derived weights vs optimal weights in a small-scale simulation setup under a range of different parameters

Two genetically correlated phenotypes were simulated under different values of h^2 (SNP heritability), r_G (genetic correlation), N (sample size) and M (number of markers). For each combination of parameters we calculated weights using Eq. [15] (red bars) and using the approximation $w_k = \boldsymbol{r}_{\boldsymbol{G}_{k,f}}\sqrt{\boldsymbol{h}_{\boldsymbol{k}}^{\boldsymbol{2}}\boldsymbol{N}_{\boldsymbol{k}}}$ for focal trait f and additional traits k, which assumes that the SNP effects of each trait k have equal variance (blue bars). For each parameter combination we replicated simulations 50 times and determined the weight ratio which resulted in the highest prediction accuracy (box plots and black dots). This confirms that the weights in Eq. [15] result in the highest prediction accuracy. The approximation may work in well in some cases, but ignores covariances among traits other than the focal trait.

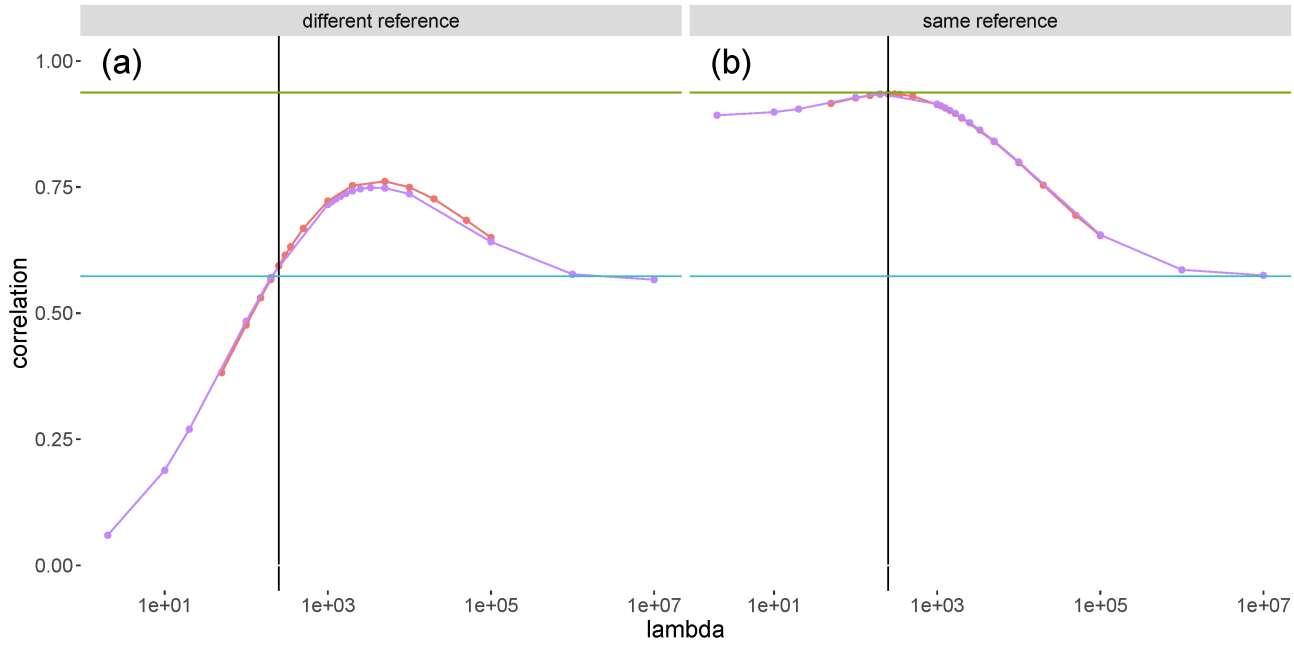

Figure 33: Comparison of the accuracy of different methods to estimate simulated SNP effects

Blue line: GWAS effect estimates (univariate OLS). Green line: BLUP estimates. Pink line: SBLUP estimates as a function of λ. Purple line: LDpred-inf estimates as a function of λ (where λ is calculated as $\frac{M}{h^2}$). Simulations are based on 20,000 individuals. 1000 SNPs from chromosome 22 were used so that the LD window included all SNPs. In (a) the external reference for SBLUP / LDpred-inf is the same dataset as the genotypes which were used for simulation. Choosing an LD reference dataset which is different from the genotypes used for simulation (b) lowers prediction accuracy and increases the optimal value of λ.

Application to Psychiatric disorders. We then applied our approach to the PGC schizophrenia [58,143] and bipolar data, two psychiatric disorders known to have a high genetic correlation [10]. The availability of combined individual-level data for both disorders enabled a direct comparison of the MT-BLUP [66] and wMT-SBLUP approaches. We calculated all predictors for the previously used [66] PGC wave 1 (PGC1) data sets [143] and compared the prediction accuracy (correlation between predicted values and phenotypes adjusted for sex, cohort and the first 20 principal components) across diseases and approaches. We find comparable but slightly lower accuracies in the wMT-SBLUP predictors as compared to the MT-BLUP predictors (0.151 vs 0.156 in bipolar disorder and 0.217 vs 0.219 in schizophrenia) and an increase in prediction accuracy as compared to the

single-trait (BLUP) predictors (0.128 in bipolar disorder, 0.198 in schizophrenia) (**Figure 34**). Our results demonstrate that creating SBLUP genetic predictors using an independent LD reference sample, and combining these in a weighted sum results in prediction accuracy comparable to a full MT-BLUP prediction for common complex disease traits, at a much lower computational burden.

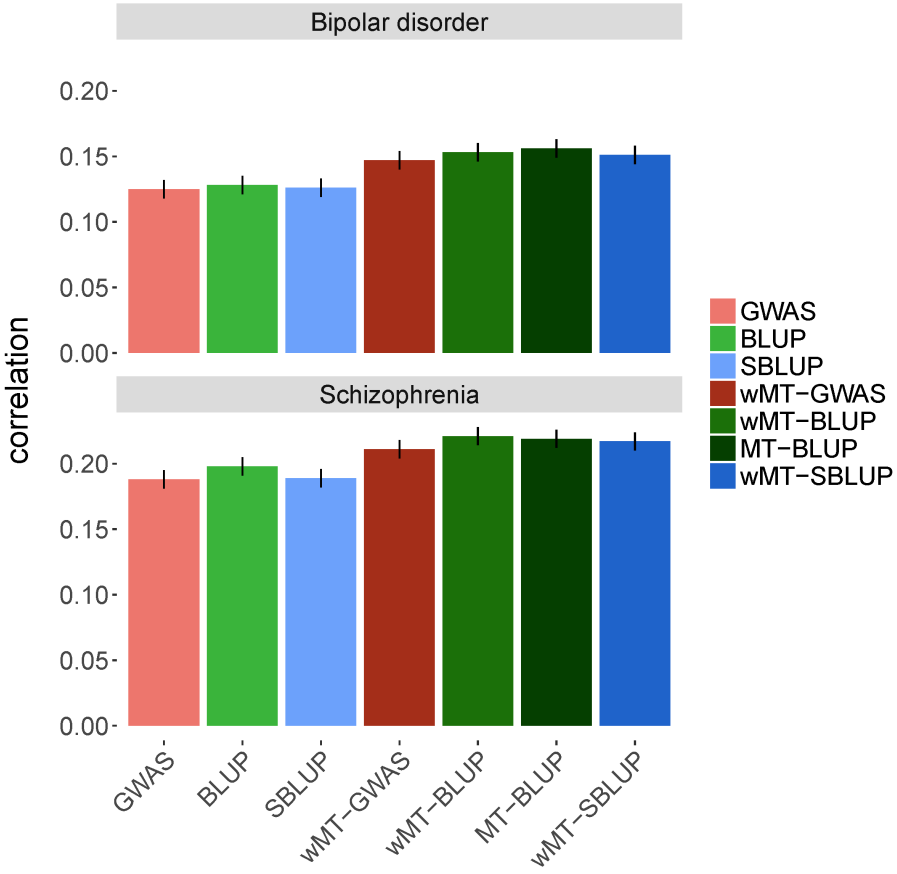

Figure 34: Prediction accuracy for schizophrenia and bipolar disorder from several single-trait and multi-trait predictors

Prediction accuracy of seven different types of predictors using PGC1 schizophrenia and bipolar disorder data. Single-trait predictor (lighter colors) are on the left, multi-trait predictors (darker colors) are on the right. Black error bars indicate correlation coefficient standard errors, calculated as $se_r = \sqrt{\frac{1-r^2}{n-2}}$.

We then applied our approach to the larger PGC wave 2 (PGC2) data sets for schizophrenia [58] and bipolar disorder (online methods), which included the PGC1 data. To test whether the addition of more cohorts improved prediction accuracy, we estimate wMT-SBLUP predictors in the PGC2 data. Having shown the resemblance of wMT-SBLUP and MT-BLUP by theory, simulation, and in the PGC1 data, we refrained from running a MT-BLUP model in the PGC2 data to avoid the computational burden of analysing the combined schizophrenia bipolar data set. For schizophrenia, there were 36 cohorts (26412 cases and 32440 controls in total) and for bipolar disorder there were 23 cohorts (18865 cases and 30460 controls in total). We conducted a cohort-wise leave-one-out cross-validation approach to examine variation in prediction accuracy across cohorts.

For schizophrenia, we find that prediction accuracy increases in 20 of the 36 cohorts of the PGC2 data when using a wMT-SBLUP predictor as compared to a SBLUP predictor (**Figure 35**). However, the median correlation (0.300 with an SBLUP predictor, and 0.304 with a wMT-SBLUP predictor) and mean correlation (0.295 with a SBLUP predictor and 0.294 with a wMT-SBLUP predictor) across the 36 PGC2 cohorts did not improve with a wMT-BLUP predictor. For bipolar disorder, we find an improvement of the wMT-SBLUP predictor over the SBLUP predictor in 17 out of 23 cohorts (**Figure 36**), with a mean correlation increase from 0.212 to 0.229 and a median correlation increase from 0.210 to 0.225. To evaluate whether this is because the weights we used for schizophrenia and bipolar disorder do not represent the mixing proportions which lead to the highest accuracy in this data set or whether other factors explain the variable results across cohorts, we created multi-trait predictors using not only weights calculated from Eq. [17], but also weights corresponding to any other mixing proportion of the two disorders (**Figure 35**, **Figure 36** and **Figure 37**). This demonstrates (i) that our calculated weights are very close to the empirically optimal weights when averaged across cohorts (**Figure 37**), (ii) that there is substantial heterogeneity across cohorts as shown by the variable prediction accuracies of single-trait and cross-trait predictors across cohorts, which is supported by previous studies [58], and (iii) that for some test set cohorts, there is no mixing proportion which will lead to a multi-trait predictor which outperforms a single-trait predictor.

The larger gain in accuracy that results from supplementing a bipolar disorder predictor with schizophrenia data compared to supplementing a schizophrenia predictor with bipolar disorder data is consistent with greater power of the schizophrenia discovery sample. We

find that for both single-trait and multi-trait predictors the SBLUP predictors outperform the OLS predictors in almost all cohorts (**Figure 35** and **Figure 36**).

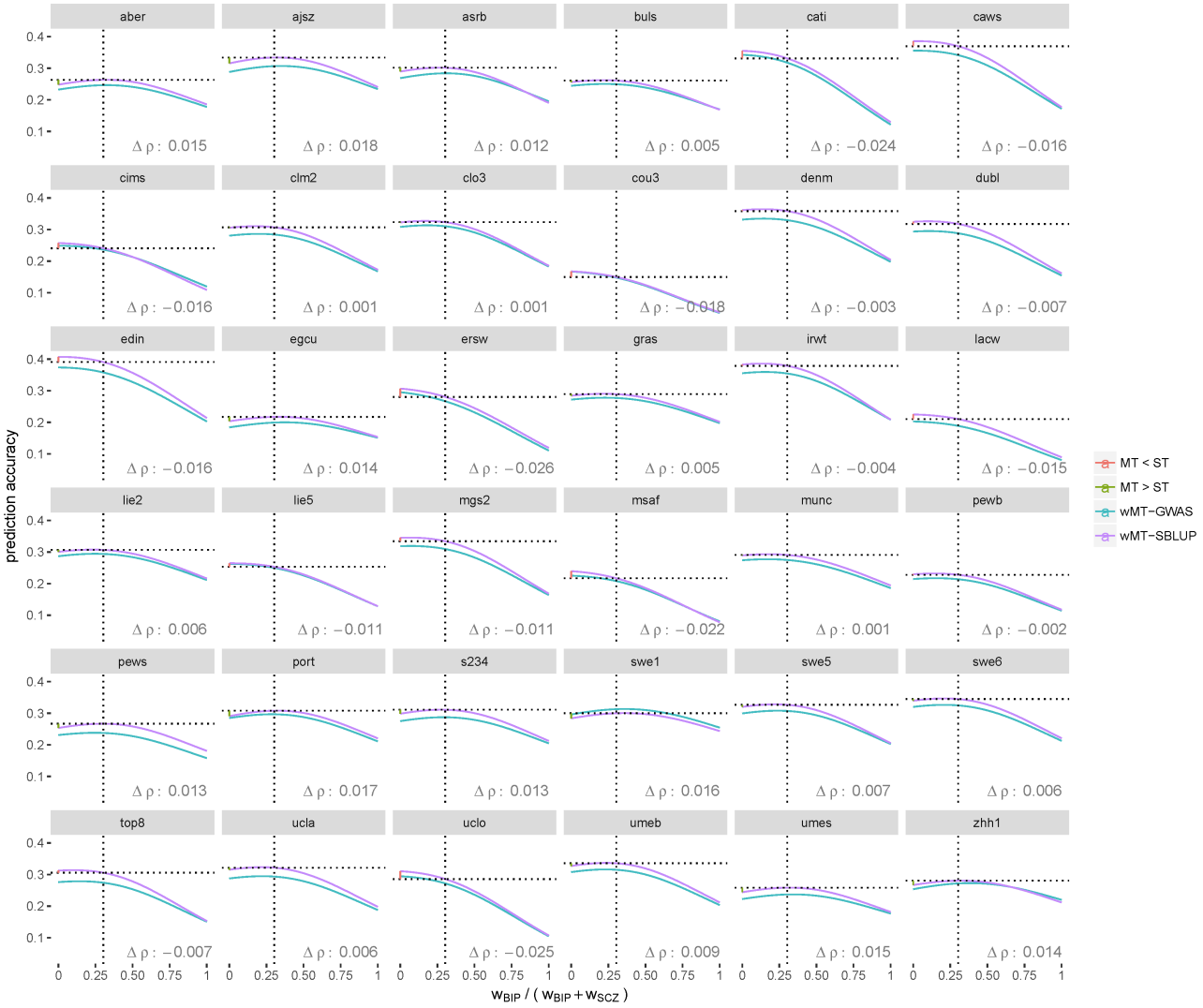

Figure 35: Prediction accuracy for schizophrenia in each schizophrenia cohort using single-trait and multi-trait, GWAS (blue) and SBLUP (purple) predictors

Each point along the x-axis represents a different multi-trait predictor with a different mixing proportion of schizophrenia and bipolar disorder data, corresponding to different weights. Dotted vertical lines indicate weights according to Eq. [15], and dotted horizontal lines indicate wMT-SBLUP prediction accuracy at these weights. If this prediction accuracy is higher than the single-trait prediction accuracy on the left hand side, the multi-trait predictor improves upon the single-trait predictor. The difference between single-trait and multi-trait accuracy (Δρ) is visualized by green or red lines and printed in each panel.

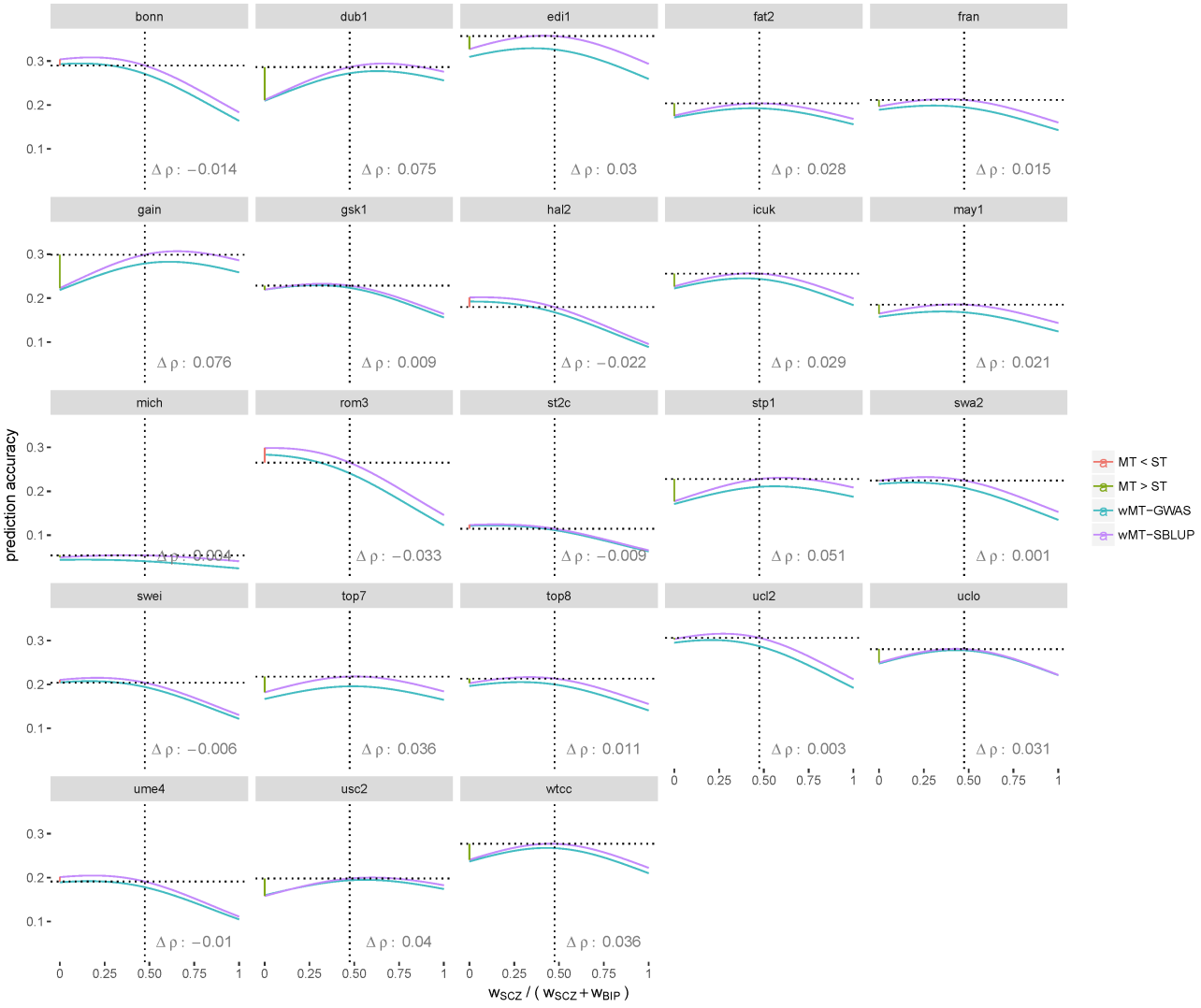

Figure 36: Prediction accuracy for bipolar in each bipolar cohort using single-trait and multi-trait, GWAS (blue) and SBLUP (purple) predictors

Each point along the x-axis represents a different multi-trait predictor with a different mixing proportion of schizophrenia and bipolar disorder data, corresponding to different weights. Dotted vertical lines indicate weights according to Eq. [15], and dotted horizontal lines indicate wMT-SBLUP prediction accuracy at these weights. If this prediction accuracy is higher than the single-trait prediction accuracy on the left hand side, the multi-trait predictor improves upon the single-trait predictor. The difference between single-trait and multi-trait accuracy (Δρ) is visualized by green or red lines and printed in each panel.

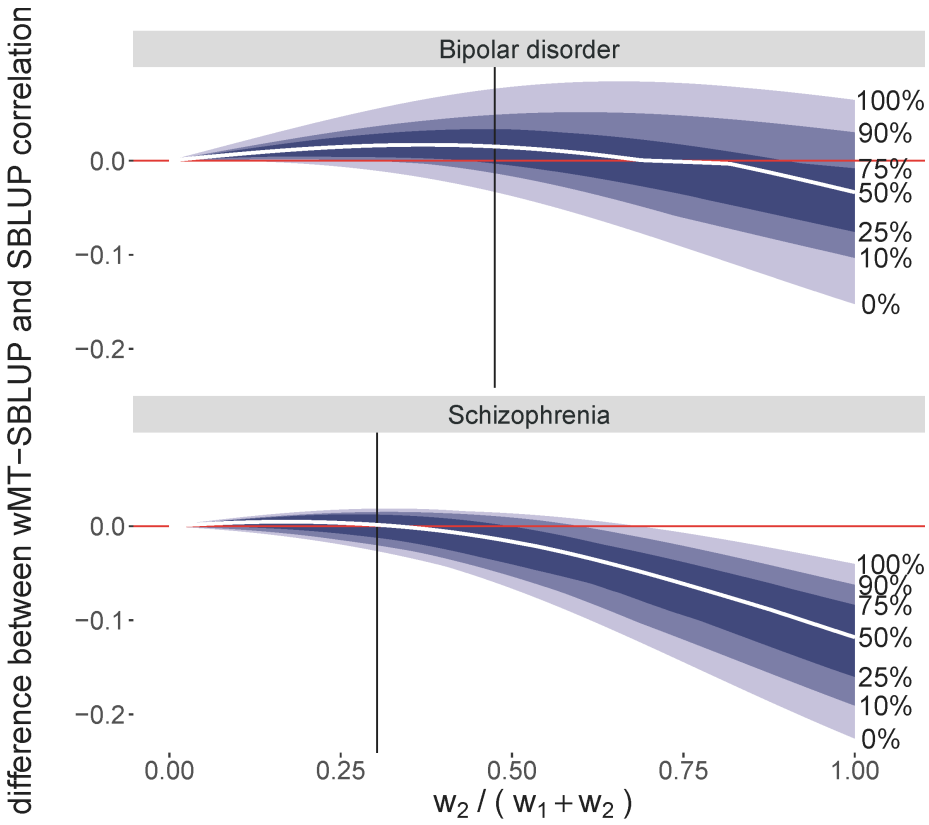

Figure 37: Prediction accuracy difference between SBLUP predictors and wMT-SBLUP predictors, summarized over all cohorts

*A summary over all cohorts shown in **Figure 35** and **Figure 36**. The y-axis now shows accuracy difference rather than absolute accuracy. For each weighting there is a distribution of prediction accuracy improvement (correlation single-trait predictor minus correlation multi-trait predictor) across all cross-validation iterations. The quantiles of this distribution are shown in shades of blue and the white line represents the median. The vertical line represents the weights derived from Eq. [15].*

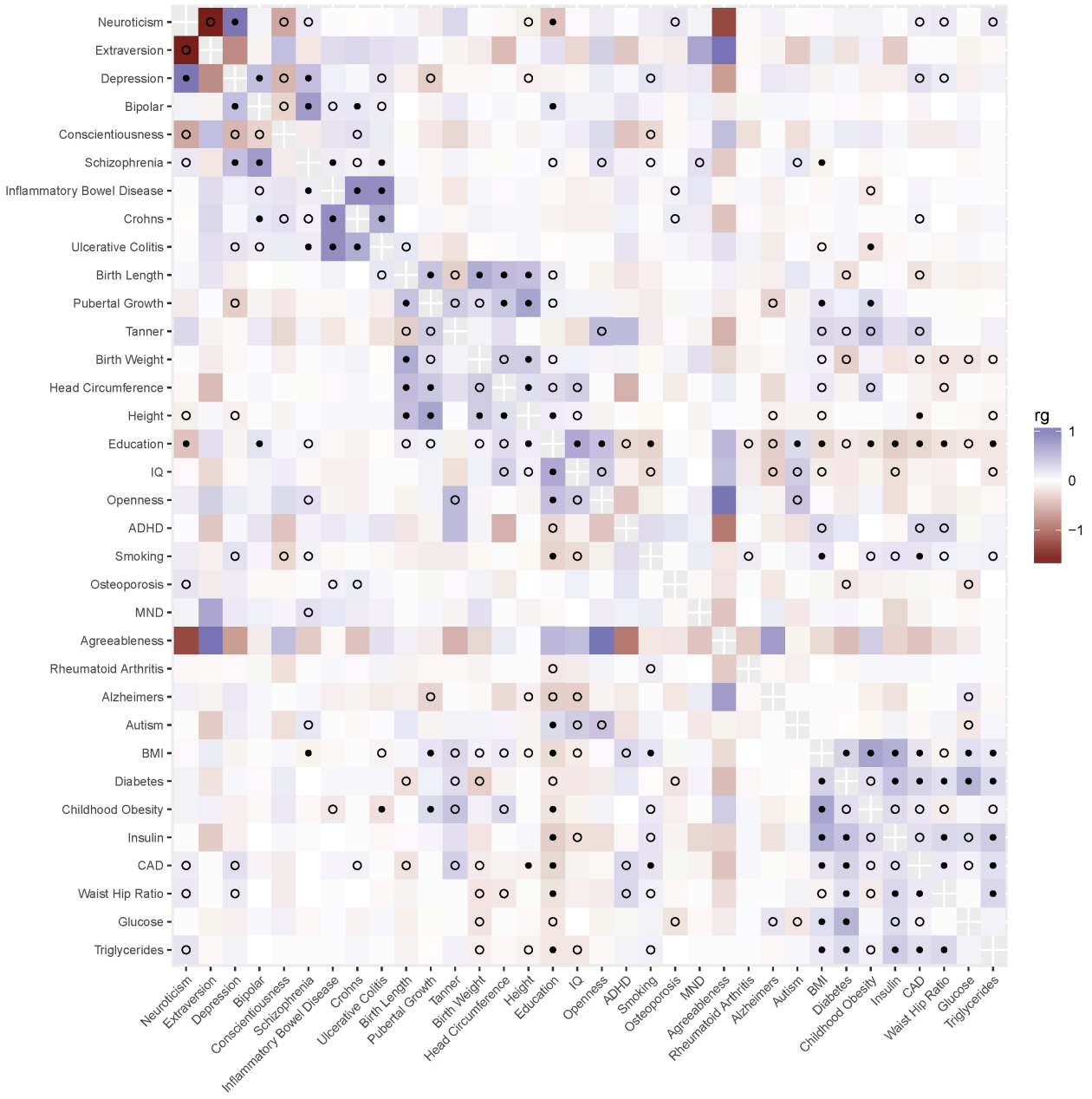

Figure 38 Genetic correlation estimates between 34 traits

LD score regression was used to estimate the genetic correlation based on summary statistics for each trait. Empty circles indicate a genetic correlation p-value lower than 0.05, filled circles indicate a genetic correlation p-value smaller than the Bonferroni threshold 0.05 / 561 = 8.91e-05. Traits are ordered according to hierarchical clustering based on the absolute value of genetic correlation estimates. Summary statistics were obtained from various sources, for details see ***Table 15****.*

Application to phenotypes recorded in a large population study. In principle any number of traits can be combined into a multi-trait predictor at almost no computational cost. We therefore extended our approach to create wMT-SBLUP predictors from 34 phenotypes for which we could access summary statistics. In order to calculate wMT-SBLUP weights, we used LD score regression to estimate SNP-heritability and genetic correlations of the 34 summary statistics traits. The results are mostly in line with previous reports [9] (**Figure 38**, **Table 17**). As test set we used 112,338 individuals in the UK Biobank data. We matched six of the 34 discovery traits to traits in the UK Biobank (**Table 15**) and created wMT-SBLUP predictors. For the wMT-SBLUP predictor of each focal trait we included predictor traits with genetic correlation p-value < 0.05. For all traits, wMT-SBLUP genetic predictors were more accurate than any single-trait (SBLUP) predictor (**Figure 39**). We observe the largest increases in accuracy for Type 2 diabetes (47.8%) and depression (34.8%). Accuracy for height (0.7%) and BMI (1.4%) increase only marginally. As shown in our theory and simulation study, the magnitude of increase in prediction accuracy of a wMT-SBLUP predictor over a single trait SBLUP predictor depends upon the prediction accuracies of all of the traits included in the index and the genetic correlation among phenotypes. As GWAS sample sizes increase and genomic predictors increase in accuracy, a wMT-SBLUP approach will likely become increasingly beneficial.

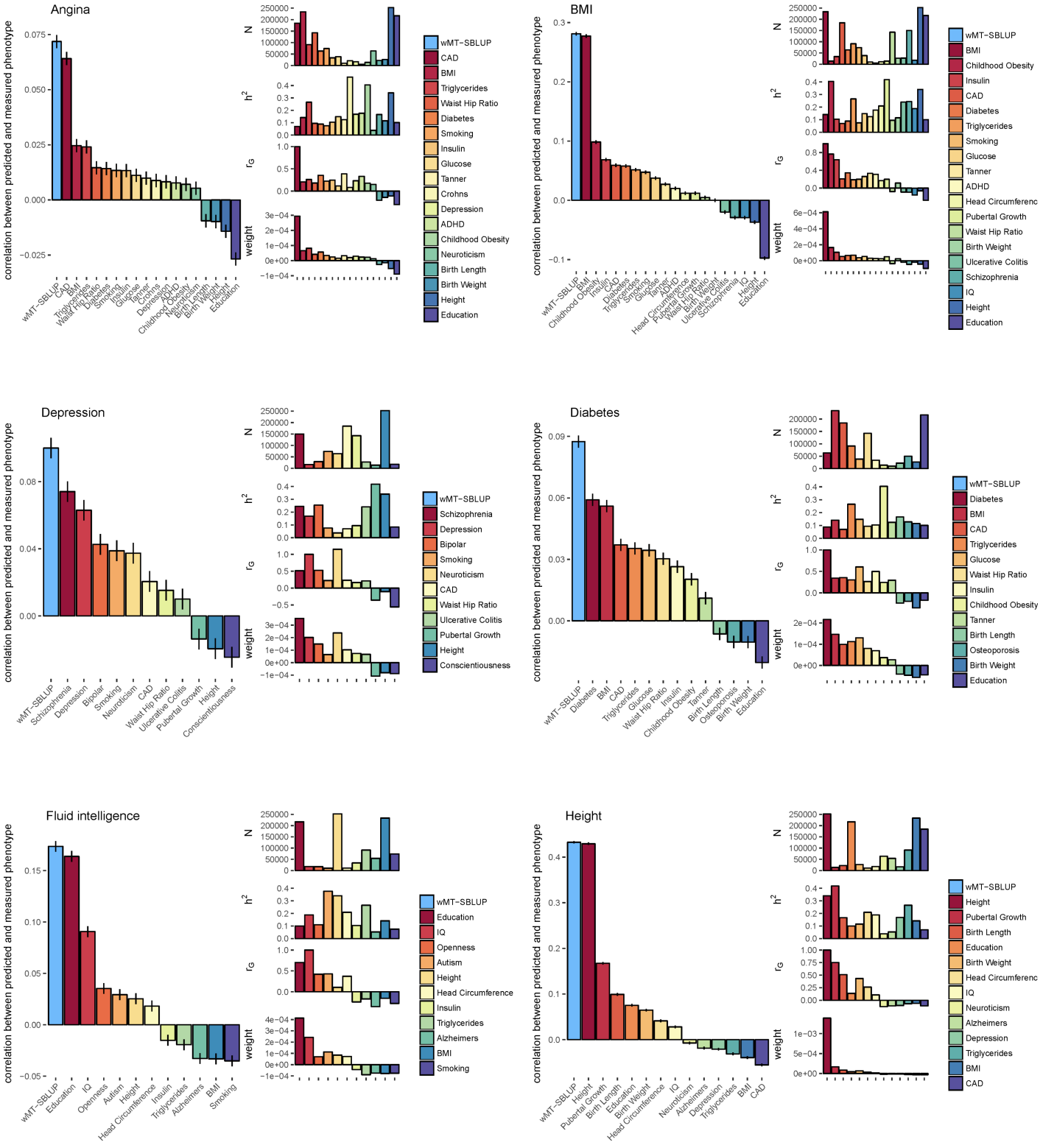

Figure 39: Prediction accuracy for single-trait and multi-trait predictors in UK Biobank traits (SBLUP)

Prediction accuracy for six traits in the UK Biobank for multi-trait predictors (light blue bars, wMT-SBLUP) and single-trait predictors (colourful bars on the right, SBLUP). Black bars

show the correlation coefficient standard error. The multi-trait predictors for each trait are composed of all traits for which colourful bars are shown (r_G p-value < 0.05). Smaller bars on the right show, from top to bottom, sample size, SNP-heritability, r_G, and weights (given by Eq. [15]) for each trait.

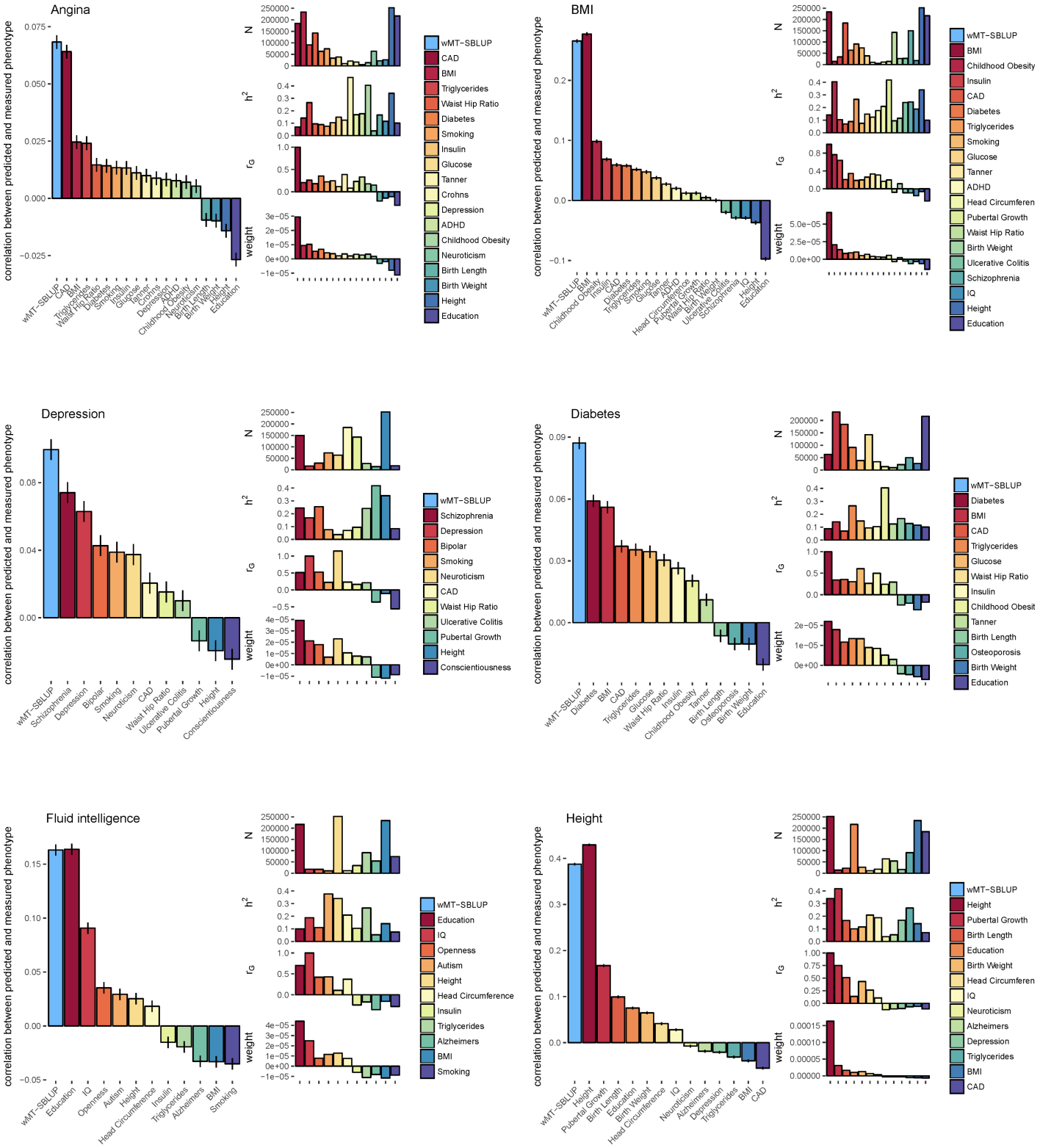

Figure 40: Prediction accuracy for single-trait and multi-trait predictors in UK Biobank traits (OLS)

Prediction accuracy for six traits in the UK Biobank for multi-trait predictors (light blue bars, wMT-OLS) and single-trait predictors (colorful bars on the right, OLS). Black bars show the correlation coefficient standard error. The multi-trait predictors for each trait are composed of all traits for which colorful bars are shown (rG p-value < 0.05). Smaller bars on the right show, from top to bottom, sample size, SNP-heritability, rG, and weights (given by Eq. [24]) for each trait.

Discussion

In summary, we demonstrate that multivariate predictors derived from GWAS summary statistics can increase prediction accuracy in a wide range of traits. This approach has particular utility in risk prediction of traits for which it is hard to generate large sample sizes for GWAS, as the increase in prediction accuracy over a standard genetic predictor is greatest when the additional traits included in the predictor have a high genetic correlation with, and are better powered than, the trait to be predicted.

Special consideration should be given to the risk of sample overlap between the summary statistics data used to create the predictor and the prediction target. Sample overlap will lead to inflated estimates of accuracy, and while here we were able to take steps to avoid individuals being recorded across multiple datasets, further work is required to negate these effects within this framework. An additional limitation of our method is that the conversion of OLS SNP effects to SBLUP SNP effects assumes that the true SNP effect sizes follow a normal distribution. However, results from other studies which have not made this assumption [69], show that it does not negatively impact prediction accuracy for the majority of traits [63,69]. Despite these limitations, current evidence suggests that genetic correlations among phenotypes are pervasive [9], sample sizes of GWAS are increasing [11], and public availability of genome-wide summary statistics is becoming the norm [4], meaning that genomic prediction of complex common disease will continually improve especially when predictors of multiple phenotypes are integrated across studies within this framework.

Methods

General model

We consider a general linear mixed model:

$$\mathbf{y} = \mathbf{Wb} + \boldsymbol{\epsilon} \quad [4]$$

where **y** is the phenotype, **W** a matrix of single nucleotide polymorphism (SNP) genotypes, where values are standardized to give the *ij*th element as: $w_{ij} = \left(x_{ij} - 2p_j\right) / \sqrt{2p_j\left(1 - p_j\right)}$, with x_{ij} the number of minor alleles (0, 1, or 2) for the *i*th individual at the *j*th SNP and p_j the minor allele frequency. **b** are the genetic effects for each SNP, and $\boldsymbol{\epsilon}$ the residual error. The dimensions of **y**, **W**, **b**, and $\boldsymbol{\epsilon}$ are dependent upon the number of phenotypes, *k*, the number of SNP markers, *M*, and the number of individuals, *N*, and are described in the sections below. We denote the distributional properties *var*(**b**) = **B**, *var*($\boldsymbol{\epsilon}$) = **R** and *var*(**y**) = **WBW**' + **R**.

For human complex diseases and quantitative phenotypes, genome-wide association studies (GWAS) have typically estimated the solutions for **b** of Eq. [1] one SNP at a time using ordinary least squares (OLS) regression [204] as:

$$\hat{\mathbf{b}}_{\mathrm{OLS}} = diag[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\mathbf{y} \quad [5]$$

where $diag[\mathbf{W}'\mathbf{W}]$ has diagonal elements $w_j{}'w_j$ and off-diagonal elements of zero. However, by analysing one SNP at a time, GWAS effect size estimates do not account for the covariance structure among SNPs and they are not unbiased in the sense that $\mathrm{E}\left[\mathrm{b}\middle|\hat{\mathrm{b}}\right] = \hat{\mathrm{b}}$ [200]. Best linear unbiased predictors (BLUP) of the SNP effects have the property $\mathrm{E}\left[\mathrm{b}\middle|\hat{\mathrm{b}}\right] = \hat{\mathrm{b}}$ are used in genomic prediction in animal and plant breeding [195], and more recently in human medical genetics, yielding improved prediction accuracy for a number of traits over genetic predictors created from OLS SNP estimates [66,168]. In a general form, BLUP solutions for **b** of Eq. [1] can be written using Henderson's mixed model equations [205] as:

$$\hat{\mathbf{b}}_{\mathrm{BLUP}} = [\mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{B}^{-1}]^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \quad [6]$$

and if **R** is diagonal, then Eq. [6] can be reduced to:

$$\hat{\mathbf{b}}_{\mathrm{BLUP}} = [\mathbf{W}'\mathbf{W} + \mathbf{B}^{-1}\mathbf{R}]^{-1}\mathbf{W}'\mathbf{y} \quad [7]$$

Below, we describe how Eq. [6] and Eq. [7] can be used to estimate BLUP SNP effects for a single trait and for multiple traits jointly, from individual-level phenotype-genotype data. We then show how Eq. [6] and [7] can be approximated to obtain BLUP SNP effects for

single and multiple traits in the absence of individual-level data, from publically available GWAS summary statistics and an independent reference sample.

Estimation of BLUP SNP effects for a single trait

For a univariate analysis of trait *k*, **y** of Eq. [4] is a column vector of length *N* x 1 and **W** has dimension *N* x *M*. Assuming **b** is an *M* x 1 vector of random SNP effects for trait *k*, with distribution $\mathbf{b} \sim N(0, \mathbf{I}_M \sigma^2_{b_k})$, then $\mathbf{B} = \mathbf{I}_M \sigma^2_{b_k}$, where $\mathbf{I}_M$ is an identity matrix of dimension *M*. $\boldsymbol{\epsilon}$ of Eq. [1] is a column vector of independent residual effects, with distribution $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_N \sigma^2_{\epsilon_k})$, giving $\mathbf{R} = \mathbf{I}_N \sigma^2_{\epsilon_k}$, where $\mathbf{I}_N$ is an identity matrix of dimension *N*. Substituting these expressions into Eq. [6] means that Eq. [7] can then be written as:

$$\hat{\mathbf{b}}_{\mathrm{BLUP}_k} = [\mathbf{W}_k' \mathbf{W}_k + \mathbf{I}_M \lambda_k]^{-1} \mathbf{W}_k' \mathbf{y}_k \qquad [8]$$

with $\lambda_k = \sigma^2_{\epsilon_k} / \sigma^2_{\mathrm{b}_k}$.

Joint estimation of BLUP SNP effects for multiple traits

When phenotypes are genetically correlated, measurements on each trait provide information on the genetic values of the other correlated traits [201,202,206]. Recent studies have shown that prediction accuracy of common complex disease can be improved by estimating SNP effects for multiple traits jointly within a multivariate mixed effects model [66,168].

If *k* traits are measured on different individuals, with N_k observations for trait *k*, the elements of Eq. [4] become: $\mathbf{y}' = [\boldsymbol{y}_1' \ldots \boldsymbol{y}_k'], \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{W}_k \end{bmatrix}$, and $\mathbf{R} = diag[\mathbf{R}_k] = diag\left[\mathbf{I}_{N_k} \sigma^2_{\epsilon_k}\right]$, a diagonal matrix of length $N = \sum_k N_k$. $\mathbf{B} = \boldsymbol{\Sigma}_b \otimes \mathbf{I}_M$, where $\boldsymbol{\Sigma}_b$ is a *k* x *k* matrix, with diagonal elements $\sigma^2_{b_k}$ and off-diagonal elements the covariances of SNP effects between traits *k* and *l*, $\sigma_{b_{k,l}}$. For Kronecker products, $\mathbf{B}^{-1} = \boldsymbol{\Sigma}_b{}^{-1} \otimes \mathbf{I}_M$ and substituting these expressions directly into Eq. [6] means that multi-trait BLUP solutions for **b** can be obtained in Eq. [7] as:

$$\hat{\mathbf{b}}_{\mathrm{MT-BLUP}} = \left[\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_b{}^{-1} \otimes \mathbf{I}_M\right]^{-1} \mathbf{W}'\mathbf{y}, \qquad [9]$$

with $\boldsymbol{\Sigma}_\epsilon = diag[\sigma^2_{\epsilon_k}]$, a diagonal *k* x *k* matrix. For a two-trait example Eq. [9] expands to:

$$\hat{\mathbf{b}}_{\mathrm{MT-BLUP}} = \left[\begin{bmatrix}\mathbf{W_1'W_1} & 0\\ 0 & \mathbf{W_2'W_2}\end{bmatrix} + \begin{bmatrix}\mathbf{I}_M\sigma^2_{\epsilon_1} & 0\\ 0 & \mathbf{I}_M\sigma^2_{\epsilon_2}\end{bmatrix}\begin{bmatrix}\mathbf{I}_M\sigma^2_{b_1} & \mathbf{I}_M\sigma_{b_{1,2}}\\ \mathbf{I}_M\sigma_{b_{2,1}} & \mathbf{I}_M\sigma^2_{b_2}\end{bmatrix}^{-1}\right]^{-1}\begin{bmatrix}\mathbf{W_1} & 0\\ 0 & \mathbf{W_2}\end{bmatrix}'\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix}$$

[10]

Estimation of BLUP SNP effects from summary statistics for multiple traits

Estimating SNP effects for multiple traits jointly in Eq. [9] requires individual-level genotype and phenotype data across a range of common complex diseases and quantitative phenotypes, which are not readily available in human medical genetics due to privacy concerns and data sharing restrictions. Additionally, Eq. [9] requires a series of computationally intensive *M* x *k* equations to be solved. However, these issues can be overcome by approximating Eq. [9] using publically available GWAS summary statistic data and an independent genomic reference sample.

Single trait approximate BLUP SNP effects can be obtained from GWAS summary statistics (SBLUP: summary statistic approximate BLUP) by replacing $\mathbf{W}_k'\mathbf{W}_k$ and $\mathbf{W}_k'\mathbf{y}_k$ of Eq. [8] by their expectation, which are $\mathbb{E}[\mathbf{W}_k'\mathbf{W}_k] = N_k\mathbf{L}$ and $\mathbb{E}[\mathbf{W}_k'\mathbf{y}_k] = N_k\hat{\mathbf{b}}_{\mathrm{OLS}_k}$ respectively, where **L** is an *M* x *M* scaled SNP linkage disequilibrium (LD) correlation matrix estimated from a reference SNP dataset and $\hat{\mathbf{b}}_{\mathrm{OLS}_k}$ are obtained from publically available GWAS summary statistics [67]. GWAS summary statistics report effect estimates of SNPs on an unstandardized scale, and not $\hat{\mathbf{b}}_{\mathrm{OLS}}$ as it is defined here. To obtain $\hat{\mathbf{b}}_{\mathrm{OLS}}$ from GWAS summary statistics, the effect of each SNP must be multiplied by the standard deviation of each SNP: $\hat{\mathbf{b}}_{\mathrm{OLS_j}} = \hat{\mathbf{b}}_{\mathrm{OLS-UNSCALED_j}} * \sqrt{2p_j(1-p_j)}$. Eq. [8] can then be written as:

$$\begin{aligned}\hat{\mathbf{b}}_{\mathrm{SBLUP}_k} &= [N_k\mathbf{L} + \mathbf{I}_M\lambda_k]^{-1}N_k\hat{\mathbf{b}}_{\mathrm{OLS}_k}\\ &= [\mathbf{L} + \mathbf{I}_M\lambda_k/N_k]^{-1}\hat{\mathbf{b}}_{\mathrm{OLS}_k}\end{aligned} \qquad [11]$$

The shrinkage parameter is $\lambda_k = \sigma^2_{\epsilon_k}/\sigma^2_{b_k} = M\sigma^2_{\epsilon_k}/h^2_{SNP_k} = M(1-h^2_{SNP_k})/h^2_{SNP_k}$, under the assumption of phenotypic variance of 1 which makes the proportion of phenotypic variance of trait *k* attributable to the SNPs $h^2_{SNP_k} = M\sigma^2_{b_k}$.

This approach was implemented in [68] and is similar to the model presented by Vilhjálmsson *et al.* [69] but with two differences. The first is that the shrinkage parameter of Vilhjálmsson *et al.* [69] is $\lambda_k = M/h^2_{SNP_k}$ as they assume that the error variance is 1 rather than $1 - h^2_{SNP_k}$ in our implementation. The second difference is that Vilhjálmsson *et al.* [69] calculate BLUP effects for blocks of a certain number of SNPs following a tiling window approach giving a block diagonal structure to L, whereas our implementation within the software GCTA (see URLs) follows a sliding window approach giving a banded diagonal to **L**. Assuming an error variance of $1 - h^2_{SNP_k}$ is more appropriate because cumulatively the SNP markers explain $h^2_{SNP_k}$ of the phenotypic variance. Additionally, a banded diagonal for **L** is also appropriate as it captures a greater extent of the long-range LD (**Figure 33**). In both implementations a window is used to capture the LD around SNP markers in order to avoid the large computational costs of inverting a dense *M* dimensional SNP LD matrix, with only little loss of information (see below).

For multiple phenotypes, the elements of Eq. [11] become: $\hat{\mathbf{b}}_{\mathrm{OLS}}{}' = \left[\hat{\mathbf{b}}_{\mathrm{OLS}_1}{}' \ldots \hat{\mathbf{b}}_{\mathrm{OLS}_k}{}'\right]$ and **N** = $\begin{bmatrix} \mathbf{N}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{N}_k \end{bmatrix}$, meaning that Eq. [11] can be extended as:

$$\hat{\mathbf{b}}_{\mathrm{MT-SBLUP}} = \left[\mathbf{I}_k \otimes \mathbf{L} + \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_b{}^{-1} \boldsymbol{N}^{-1} \otimes \mathbf{I}_M\right]^{-1} \hat{\mathbf{b}}_{\mathrm{OLS}} \qquad [12]$$

Eq. [12] approximates Eq. [9] using only publically available GWAS summary statistic data and an independent genomic reference sample. However, there remains the large computational cost associated with the inversion of the non-diagonal matrix $\left[\mathbf{I}_k \otimes \mathbf{L} + \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_b{}^{-1} \boldsymbol{N}^{-1} \otimes \mathbf{I}_M\right]$.

Index weighted multi-trait BLUP SNP effects from summary statistics

An alternative to Eq. [12], is to obtain *k* $\hat{\mathbf{b}}_{\mathrm{MT-SBLUP}}$ effects by combining together *k* single trait $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ estimates of Eq. [11], using an optimal index weighting for each trait. The index weighting to derive $\hat{\mathbf{b}}_{\mathrm{MT-SBLUP}}$ from $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ estimates is identical to the index weighting to derive $\hat{\mathbf{b}}_{\mathrm{MT-BLUP}}$ from $\hat{\mathbf{b}}_{\mathrm{BLUP}}$ estimates.

For SNP *j*, and focal trait *f*, we have $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ values for *k* traits, and we wish to obtain the index weights, $\mathrm{w}_{j,k}$, for each $\hat{\mathbf{b}}_{\mathrm{SBLUP}_{j,k}}$effect as:

$$\hat{\mathbf{b}}_{\mathrm{wMT-SBLUP}_{j,f}} = \sum_k \mathrm{w}_{\mathrm{SBLUP},j,k} \hat{\mathrm{b}}_{\mathrm{SBLUP}_{j,k}} = \mathbf{w}'_{\mathrm{SBLUP},j} \hat{\mathbf{b}}_{\mathrm{SBLUP}_j} \qquad [13]$$

In animal and plant breeding, selection indices have been developed, which combine many single trait BLUP predictors of an individual's genetic value together in an index weighting to optimise the selection of individuals with the most favourable multi-trait phenotype for breeding programs [207–210]. Utilising a selection index approach, the solution for $\mathbf{w}_{\mathrm{SBLUP}}$ of Eq. [13] can be obtained as:

$$\mathbf{w}_{\mathrm{SBLUP}} = \mathbf{V}_{\mathrm{SBLUP}}^{-1}\mathbf{C}_{\mathrm{SBLUP}} \quad [14]$$

where $\mathbf{C}_{\mathrm{SBLUP}}$ a *k* x 1 column vector of the covariance of the $\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}$ values of the k traits, with the true genetic effects of the SNPs for the focal trait, and $\mathbf{V}_{\mathrm{SBLUP}}$ a *k* x *k* variance-covariance matrix of the $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ effects:

$$\mathbf{w}_{\mathrm{SBLUP}} = \mathbf{V}_{\mathrm{SBLUP}}^{-1}\mathbf{C}_{\mathrm{SBLUP}} = \begin{bmatrix} var(\hat{\mathbf{b}}_{\mathrm{SBLUP}_1}) & \cdots & cov(\hat{\mathbf{b}}_{\mathrm{SBLUP}_1}, \hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) \\ \vdots & \ddots & \vdots \\ cov(\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}, \hat{\mathbf{b}}_{\mathrm{SBLUP}_1}) & \cdots & var(\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) \end{bmatrix}^{-1} \begin{bmatrix} cov(\mathbf{b}_{\mathrm{f}}, \hat{\mathbf{b}}_{\mathrm{SBLUP}_1}) \\ \vdots \\ cov(\mathbf{b}_{\mathrm{f}}, \hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) \end{bmatrix} \quad [15]$$

Therefore, if $\mathbf{V}_{\mathrm{SBLUP}}$ and $\mathbf{C}_{\mathrm{SBLUP}}$ can be approximated then $\hat{\mathbf{b}}_{\mathrm{MT-SBLUP}}$ of Eq. [12] can be obtained from *k* single trait $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ estimates from Eq. [11].

To derive the approximations, we first consider the diagonal elements of $\mathbf{V}_{\mathrm{SBLUP}}$ which comprise the variance of the SBLUP SNP solutions, $var(\hat{\mathbf{b}}_{\mathrm{SBLUP}_k})$. These can be approximated from theory under the assumption that $\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}$ have BLUP properties $\mathrm{E}[\mathrm{b}|\hat{\mathrm{b}}] = \hat{\mathrm{b}}$, which in turn implies that $cov(\mathbf{b}_k, \hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) = var(\hat{\mathbf{b}}_{\mathrm{SBLUP}_k})$. Following Daetwyler *et al*. [169] and Wray *et al*. [211], the squared correlation between a phenotype, $\boldsymbol{y}_k$, in an independent sample and a single trait BLUP predictor of the phenotype, $\hat{\mathbf{g}}_{\mathrm{BLUP}_k}$, is approximately:

$$R^2_{\boldsymbol{y}_k, \hat{\mathbf{g}}_{\mathrm{BLUP}_k}} = R^2_k \approx h^2_k / \left(1 + M_{eff}\,(1 - R^2_k)/(N_k h^2_k)\right) \quad [16]$$

where $\hat{\mathbf{g}}_{\mathrm{BLUP}_k} = \mathbf{W}\hat{\mathbf{b}}_{\mathrm{BLUP}_k}$ and h^2_k is the proportion of phenotypic variance attributable to additive genetic effects for trait *k*. Note that M_{eff} is the effective number of chromosome segments or the number of independent SNPs which is a function of effective population size (N_e) and can be empirically obtained as an inverse of the variance of genomic relationships [212,213]. Here, we use an estimate of M_{eff} of 60,000, which is in line both with our estimates from the genomic relationships in our simulation data and with previously reported estimates [33]. With a phenotypic variance of 1 and individual-level genetic effects

$\mathbf{g}_k = \mathbf{W}\mathbf{b}_k$, then $h_k^2 = \sigma_{g_k}^2 = M\sigma_{b_k}^2$ and the squared correlation between the true, $\mathbf{g}_k$, and estimated BLUP effects, $\hat{\mathbf{g}}_{\mathrm{BLUP}_k}$, is:

$$R^2_{\mathbf{g}_k,\hat{\mathbf{g}}_{\mathrm{BLUP}_k}} = R_k^2/h_k^2 \qquad [17]$$

rearranging Eq. [17] gives $R_k^2 = h_k^2 R^2_{\mathbf{g}_k,\hat{\mathbf{g}}_{\mathrm{BLUP}_k}} = h_k^2 \frac{cov\left(\mathbf{g}_k,\hat{\mathbf{g}}_{\mathrm{BLUP}_k}\right)^2}{var(\mathbf{g}_k)var(\hat{\mathbf{g}}_{\mathrm{BLUP}_k})}$, which given the BLUP properties $cov\left(\mathbf{g}_k,\hat{\mathbf{g}}_{\mathrm{BLUP}_k}\right) = var(\hat{\mathbf{g}}_{\mathrm{BLUP}_k})$and $h_k^2 = \sigma_{g_k}^2$with a phenotypic variance of 1, reduces to $R_k^2 = cov\left(\mathbf{g}_k,\hat{\mathbf{g}}_{\mathrm{BLUP}_k}\right) = var(\hat{\mathbf{g}}_{\mathrm{BLUP}_k}) = Mvar(\hat{\mathbf{b}}_{\mathrm{BLUP}_k})$. Therefore:

$$var(\hat{\mathbf{b}}_{\mathrm{BLUP}_k}) = \frac{var(\hat{\mathbf{g}}_{\mathrm{BLUP}_k})}{M} = \frac{R_k^2}{M} \qquad [18]$$

Second, we consider the off-diagonal elements of $\mathbf{V}_{\mathrm{SBLUP}}$, which are comprised of the covariance of BLUP SNP solutions among the *k* traits. These can again be approximated from theory given the covariance of genetic effects among traits *k* and *l* is $cov(\mathbf{b}_k,\mathbf{b}_l) = \mathrm{r}_G h_k h_l/M$, with r_G the genetic correlation, and given the squared correlation between the true genetic effects of the SNPs, $\mathbf{b}_k$, and $\hat{\mathbf{b}}_{\mathrm{BLUP}_k}$ which is given by Eq. [17] as $R^2_{\mathbf{b}_k,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}} = \frac{R_k^2}{M}/\frac{h_k^2}{M} = R_k^2/h_k^2$. The covariance of BLUP SNP predictors is then:

$$cov(\hat{\mathbf{b}}_{BLUP\,k},\hat{\mathbf{b}}_{BLUP_l}) = \frac{R_k^2}{h_k^2}\cdot\frac{R_l^2}{h_l^2}cov(\mathbf{b}_k,\mathbf{b}_l) = \frac{\mathrm{r}_G R_k^2 R_l^2}{h_k h_l M} \qquad [19]$$

Finally, we can consider the column vector $\mathbf{C}_{\mathrm{SBLUP}}$, which is composed of the covariance between the true genetic effects of the SNPs for the focal trait, $\mathbf{b}_f$, and $\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}$for all of the *k* traits. The first element of $\mathbf{C}_{\mathrm{SBLUP}}$ is covariance between the true genetic effects of the SNPs for the focal trait $\mathbf{b}_f$and $\hat{\mathbf{b}}_{\mathrm{SBLUP}_f}$for the focal trait $cov(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{BLUP}_f}) = var(\hat{\mathbf{b}}_{\mathrm{BLUP}_f}) = \frac{R_f^2}{M}$. The remaining elements of $\mathbf{C}_{\mathrm{SBLUP}}$ are $cov(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{BLUP}_k})$which can be approximated from theory by considering a regression of $\mathbf{b}_f$ on $\mathbf{b}_k$ where the regression coefficient $\beta_{f,k} = \mathrm{r}_G\sqrt{var(\mathbf{b}_f)/var(\mathbf{b}_k)}$. The covariance of $\mathbf{b}_f$and $\hat{\mathbf{b}}_{\mathrm{BLUP}_k}$can then be written as:

$$cov(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}) = cov(\beta_{f,k}\mathbf{b}_k,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}) = \mathrm{r}_G\frac{R_k^2}{M}\cdot\frac{h_f}{h_k} \qquad [20]$$

If we consider a two-trait example where the focal trait that we want to predict is matched to the first of the two traits, Eq. [18-20] combine as:

$$\mathbf{w}_{\mathrm{SBLUP}} = \mathbf{V}_{\mathrm{SBLUP}}^{-1}\mathbf{C}_{\mathrm{SBLUP}} = \begin{bmatrix} \frac{R_1^2}{M} & \frac{\mathrm{r}_G R_1^2 R_2^2}{h_1 h_2 M} \\ \frac{\mathrm{r}_G R_1^2 R_2^2}{h_1 h_2 M} & \frac{R_2^2}{M} \end{bmatrix}^{-1}\begin{bmatrix} \frac{R_1^2}{M} \\ \mathrm{r}_G\frac{R_2^2}{M}\cdot\frac{h_1}{h_2} \end{bmatrix} \qquad [21]$$

giving the index for the focal trait as: $\hat{\mathbf{b}}_{\mathrm{wMT-SBLUP}_f} = \mathrm{w}_1\hat{\mathbf{b}}_{SBLUP_f} + \mathrm{w}_2\hat{\mathbf{b}}_{\mathrm{SBLUP}_2}$with solutions for the index weights of:

$$w_f = \left(1 - \frac{{r_G}^2 R_2^2}{h_2^2}\right) \Big/ \left(1 - \frac{{r_G}^2 R_f^2 R_2^2}{h_f^2 h_2^2 M}\right)$$

$$= \left(1 - {r_G}^2 R^2_{\mathbf{b}_2, \hat{\mathbf{b}}_{\mathrm{BLUP}_2}}\right) \Big/ \left(1 - {r_G}^2 R^2_{\mathbf{b}_f, \hat{\mathbf{b}}_{\mathrm{BLUP}_f}} R^2_{\mathbf{b}_2, \hat{\mathbf{b}}_{\mathrm{BLUP}_2}}\right), \text{ and}$$

$$w_2 = \left(r_G / \left(h_f h_2\right)\left(h_f^2 - R_1^2\right)\right) / \left(1 - \frac{{r_G}^2 R_f^2 R_2^2}{h_f^2 h_2^2 M}\right)$$

$$= r_G \left(h_f / h_2\right) \left(1 - R^2_{\mathbf{b}_f, \hat{\mathbf{b}}_{\mathrm{BLUP}_f}}\right) \Big/ \left(1 - {r_G}^2 R^2_{\mathbf{b}_f, \hat{\mathbf{b}}_{\mathrm{BLUP}_f}} R^2_{\mathbf{b}_2, \hat{\mathbf{b}}_{\mathrm{BLUP}_2}}\right) \qquad [22]$$

For traits with low power R_k^2 is usually very small. In that case, $\mathbf{V}_{\mathrm{SBLUP}}$ can be well approximated by a diagonal matrix with entries $\frac{R_k^2}{M}$. w_f will become 1 and $w_k$for all other traits will be $r_{G_{f,k}} \frac{h_f}{h_k}$. It may appear surprising that traits with higher SNP-heritability have smaller weights than traits with lower SNP-heritability. This can be explained by the fact that the variance of each BLUP predictor (R_k^2) is approximately proportional to $h_k^4 N$ if M_{eff} is large, and thus a trait with higher SNP-heritability will still have a larger contribution to the multi-trait predictor than a trait with lower SNP-heritability.

Eq. [17] implies $R^2_{\mathbf{b}_k, \hat{\mathbf{b}}_{\mathrm{BLUP}_k}} = R^2_{\mathbf{g}_k, \hat{\mathbf{g}}_{\mathrm{BLUP}_k}} = R_k^2 / h_k^2$ and thus the index weights of Eq. [15] can be applied equally to BLUP solutions for the SNP effects, or BLUP predictors for individuals of each trait as described in the main text in Eq. [1] through [3]. Both $r_{G_{k,l}}$ and h_k^2 of Eq. [15] can be obtained from summary statistic data using LD Score regression [32] and therefore $\hat{\mathbf{b}}_{\mathrm{MT-BLUP}}$ effects of Eq. [10], which would traditionally require individual-level phenotype-genotype data for all traits, can be approximated accurately in a computationally efficient manner using only publically available GWAS summary statistic data and an independent genomic reference sample.

Index weighted multi-trait OLS SNP effects from summary statistics

In the previous section we have shown how $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ estimates for multiple traits can be combined to yield more accurate $\hat{\mathbf{b}}_{\mathrm{wMT-SBLUP}}$ SNP effects, which can be turned into $\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}}$ individual predictors that approach $\hat{\mathbf{g}}_{\mathrm{MT-BLUP}}$ accuracy. However, using a similar weighting we can also combine $\hat{\mathbf{b}}_{\mathrm{OLS}}$ estimates for multiple traits into $\hat{\mathbf{b}}_{\mathrm{wMT-OLS}}$.

For SNP *j*, and focal trait *f*, we have $\hat{\mathbf{b}}_{\mathrm{OLS}}$ values for *k* traits, and we wish to obtain the index weights, $\mathrm{w}_{j,k}$, for each $\hat{\mathbf{b}}_{\mathrm{OLS}_{j,k}}$effect as:

$$\hat{\mathbf{b}}_{\mathrm{wMT-OLS}_{j,f}} = \sum_k \mathrm{w}_{j,k}\hat{\mathrm{b}}_{\mathrm{OLS}_{j,k}} = \mathbf{w}_j'\hat{\mathbf{b}}_{\mathrm{OLS}_j} \quad [23]$$

Just like before, the optimal weights can be derived as: $\mathbf{w}_{\mathrm{OLS}} = \mathbf{V}_{\mathrm{OLS}}^{-1}\mathbf{C}_{\mathrm{OLS}}$, where $\mathbf{C}_{\mathrm{OLS}}$ is now a *k* x 1 column vector of the covariances of the $\hat{\mathbf{b}}_{\mathrm{OLS}_k}$ values of the k traits with the true genetic effects of the SNPs for the focal trait, and $\mathbf{V}_{\mathrm{OLS}}$ is a *k* x *k* variance-covariance matrix of the $\hat{\mathbf{b}}_{\mathrm{OLS}}$ effects:

$$\mathbf{w}_{\mathrm{OLS}} = \mathbf{V}_{\mathrm{OLS}}^{-1}\mathbf{C}_{\mathrm{OLS}} = \begin{bmatrix} var(\hat{\mathbf{b}}_{\mathrm{OLS}_1}) & \cdots & cov(\hat{\mathbf{b}}_{\mathrm{OLS}_1}, \hat{\mathbf{b}}_{\mathrm{OLS}_k}) \\ \vdots & \ddots & \vdots \\ cov(\hat{\mathbf{b}}_{\mathrm{OLS}_k}, \hat{\mathbf{b}}_{\mathrm{OLS}_1}) & \cdots & var(\hat{\mathbf{b}}_{\mathrm{OLS}_k}) \end{bmatrix}^{-1} \begin{bmatrix} cov(\mathbf{b}_{\mathrm{f}}, \hat{\mathbf{b}}_{\mathrm{OLS}_1}) \\ \vdots \\ cov(\mathbf{b}_{\mathrm{f}}, \hat{\mathbf{b}}_{\mathrm{OLS}_k}) \end{bmatrix}$$

[24]

The diagonal elements of $\mathbf{V}_{\mathrm{OLS}}$ are:

$$var(\hat{\mathbf{b}}_{\mathrm{OLS}_k}) = \frac{h_k^2}{M} + \frac{1}{N_k} \quad [25]$$

The off-diagonal elements for trait *k* and *l* are

$$cov(\hat{\mathbf{b}}_{\mathrm{OLS}_k}, \hat{\mathbf{b}}_{\mathrm{OLS}_l}) = \frac{\mathrm{r}_G h_k h_l}{M}. \quad [26]$$

$\mathbf{C}_{\mathrm{OLS}}$ now has elements

$$cov(\mathbf{b}_{\mathrm{k}}, \hat{\mathbf{b}}_{\mathrm{OLS}_k}) = \frac{\mathrm{r}_G h_k h_l}{M} \quad [27]$$

If we again consider a two-trait example, Eq. [25-27] combine as:

$$\mathbf{w}_{\mathrm{OLS}} = \mathbf{V}_{\mathrm{OLS}}^{-1}\mathbf{C}_{\mathrm{OLS}} = \begin{bmatrix} \frac{h_1^2}{M} + \frac{1}{N_1} & \frac{\mathrm{r}_G h_1 h_2}{M} \\ \frac{\mathrm{r}_G h_1 h_2}{M} & \frac{h_2^2}{M} + \frac{1}{N_2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{h_1^2}{M} \\ \frac{\mathrm{r}_G h_1 h_2}{M} \end{bmatrix} \quad [28]$$

These weights are considerably different from the BLUP weights, which reflects the different variances of BLUP effects and OLS effects. Here, we include this section for completeness but focus our analyses on multi-trait BLUP effects, because they are more accurate in expectation than multi-trait OLS effects.

Prediction accuracy of an index weighted multi-trait BLUP predictor

The prediction accuracy of $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}$ effects obtained from Eq. [15] can be derived by considering the correlation of $\mathbf{b}_f$and $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_k}$as:

$$r_{\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}} = \frac{cov\left(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}\right)}{\sqrt{var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}\right)var(\mathbf{b}_f)}} \quad [29]$$

Eq. [13] gives $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f} = \mathbf{w}'\hat{\mathbf{b}}_{\mathrm{BLUP}}$ and thus the covariance of $\mathbf{b}_f$ and $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}$ is:

$$cov\left(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}\right) = cov\left(\mathbf{b}_f,\mathbf{w}'\hat{\mathbf{b}}_{\mathrm{BLUP}}\right) = \mathbf{w}'cov\left(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{BLUP}}\right) = \mathbf{w}'\mathbf{C} \quad [30]$$

The variance of the $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}$ effects obtained from Eq. [15] is:

$$var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right) = var\left(\mathbf{w}'\hat{\mathbf{b}}_{\mathrm{BLUP}_k}\right) = \mathbf{w}'var\left(\hat{\mathbf{b}}_{\mathrm{BLUP}_k}\right)\mathbf{w} = \mathbf{w}'\mathbf{V}\mathbf{w} \quad [31]$$

Additionally, $\mathbf{w} = \mathbf{V}^{-1}\mathbf{C}$ and $\mathbf{V}\mathbf{w} = \mathbf{C}$, and thus $\mathbf{w}'\mathbf{C} = \mathbf{w}'\mathbf{V}\mathbf{w}$ or written another way $cov\left(\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}\right) = var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right)$following BLUP properties. Substituting into Eq. [19] the correlation of $\mathbf{b}_f$ and $\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_k}$ can then be written as:

$$r_{\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}} = var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right)/\sqrt{var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right)var\left(\mathbf{b}_f\right)}$$

$$= \sqrt{var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right)/var\left(\mathbf{b}_f\right)} \quad [32]$$

which gives the squared correlation as $R^2_{\mathbf{b}_f,\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}_f}} = var\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right)/var\left(\mathbf{b}_f\right) = \frac{R_f^2}{M}/\frac{h_k^2}{M} = R_f^2/h_k^2$**.** Therefore, the squared correlation between a phenotype, and a multiple trait index weighted BLUP predictor of the phenotype is approximately:

$$R^2_{\mathbf{y}_k,\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}_k}} = Mvar\left(\hat{\mathbf{b}}_{\mathrm{wMT-BLUP}}\right) = M\mathbf{w}'\mathbf{V}\mathbf{w}. \quad [33]$$

If we consider a two-trait example then prediction accuracy for a focal trait $R^2_{\mathbf{y}_f,\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}_k}}$ can be written as:

$$R^2_{\mathbf{y}_f,\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}_f}} = \mathbf{w}_f^2R^2_{\mathbf{y}_f,\hat{\mathbf{g}}_{\mathrm{BLUP}_f}} + \mathbf{w}_2^2R^2_{\mathbf{y}_2,\hat{\mathbf{g}}_{\mathrm{BLUP}_2}} + 2\mathrm{w}_f\mathrm{w}_2\mathrm{V}_{1,2} \quad [34]$$

where $\mathrm{V}_{1,2}$ is the off-diagonal element of the matrix $\mathbf{V}$ of Eq. [15] and [21]. The value of $R^2_{\mathbf{y}_f,\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}_f}}$ can then be compared to the prediction accuracy of the single trait BLUP predictor of Eq. [16] and to the prediction accuracy of a cross-trait predictor [57], where a BLUP predictor of the second trait is used to predict the focal trait phenotype, which is given by: $R^2_{\mathbf{y}_f,\hat{\mathbf{g}}_{\mathrm{BLUP}_2}} = R^2_{\mathbf{y}_2,\hat{\mathbf{g}}_{\mathrm{BLUP}_2}}\mathrm{r}_G\sqrt{\left(h_2/h_f\right)}$. This comparison is of interest, because we expect the multi-trait predictor to be more accurate than any available single-trait predictor, even if the most accurate single-trait predictor is across two different traits. Cross-trait prediction is equivalent to the proxy-phenotype method, which has been used to predict cognitive performance from educational attainment GWAS data [214].

Loss of prediction accuracy when approximating a BLUP predictor

Eq. [16-34] assume that $cov(\mathbf{b}_k, \hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) = var(\hat{\mathbf{b}}_{\mathrm{SBLUP}_k}) = var(\hat{\mathbf{b}}_{\mathrm{BLUP}_k})$, or in other words that SBLUP SNP solutions have BLUP properties. The use of an independent LD reference sample to create an approximate single trait BLUP predictor in Eq. [11] does not affect the covariance between the true SNP effect sizes and the approximate BLUP SNP solution, meaning that the approximate single trait BLUP predictors have BLUP properties. However, the variance of $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ is likely affected, which may potentially result in a loss of prediction accuracy of a weighted multi-trait BLUP predictor. The variance of $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ is:

$$\sigma^2_{\hat{\mathbf{b}}_{\mathrm{SBLUP}}} = [[N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}\mathbf{W}'][\mathbf{W}'\mathbf{W}\sigma^2_{\mathrm{b}} + \mathbf{I}\sigma^2_{\mathrm{e}}][\mathbf{W}[N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}]$$
$$= [N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}[(\mathbf{W}'\mathbf{W})(\mathbf{W}'\mathbf{W})\sigma^2_{\mathrm{b}} + \mathbf{W}'\mathbf{W}\sigma^2_{\mathrm{e}}][N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}$$
$$= [([N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}(\mathbf{W}'\mathbf{W})(\mathbf{W}'\mathbf{W})[N\mathbf{L} + \mathbf{I}_M\lambda]^{-1} + [N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}\mathbf{W}'\mathbf{W}\lambda_k)\ [N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}]\sigma^2_{\mathrm{b}} \quad [35]$$

The loss of information from using an independent data set as an LD reference to obtain **L**, rather than directly using the individual-level data to calculate $\mathbf{W}'\mathbf{W}$, can be approximated by considering the scenario where SNP markers are unlinked, resulting in $diag[\mathbf{L}]$. The diagonal elements of $\sigma^2_{\hat{\mathbf{b}}_{\mathrm{SBLUP}_{jj}}}$ for SNP *j* are then:

$$\sigma^2_{\hat{\mathbf{b}}_{\mathrm{SBLUP}_{jj}}} = ([N + \lambda]^{-2}diag[(\mathbf{W}'\mathbf{W})(\mathbf{W}'\mathbf{W})] + N\lambda\,[N + \lambda]^{-2})\sigma^2_{\mathrm{b}} \quad [36]$$

The diagonal elements of $diag[(\mathbf{W}'\mathbf{W})(\mathbf{W}'\mathbf{W})]$ can be approximated as $diag[(\mathbf{W}'\mathbf{W})(\mathbf{W}'\mathbf{W})] \approx N^2(1 + \mathbb{E}[r^2]M) = N^2(1 + M/N)$, where the expectation of the LD correlation of the SNPs, $\mathbb{E}[r^2]$, is 1/N as the SNP markers are unlinked. Eq. [36] can then be written as:

$$\sigma^2_{\hat{\mathbf{b}}_{\mathrm{SBLUP}_{jj}}} = ((N^2 + NM + N\lambda)/(N + \lambda)^2)\sigma^2_{\mathrm{b}}$$
$$= \sigma^2_{\mathrm{b}}N/(N + \lambda) + \sigma^2_{\mathrm{b}}NM/(N + \lambda)^2 \quad [37]$$

From Eq. [37] the squared correlation between true SNP effects and SBLUP SNP effects can be written as:

$$R^2_{\boldsymbol{b},\hat{\mathbf{b}}_{\mathrm{SBLUP}}} = N/(N + M + \lambda) = N/(N + M/h^2) \quad [38]$$

This can be contrasted to Eq. [17], which gives the squared correlation between the true genetic effects of the SNPs, $\mathbf{b}_k$, and $\hat{\mathbf{b}}_{\mathrm{BLUP}_k}$ as:

$$R^2_{\mathbf{b}_k,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}} = \frac{R^2_k}{M}\Big/\frac{h^2_k}{M} = R^2_k/h^2_k = 1/(1 + M\,(1 - R^2_k)/(N_k h^2_k))$$
$$= N_k/(N_k + M\,(1 - R^2_k)/h^2_k) \quad [39]$$

Eq. [39] is similar to Eq. [38] apart from the factor $1 - R_k^2$. Therefore, the relative loss of prediction accuracy from using an SBLUP predictor is given as a ratio of Eq. [39] and Eq. [38] as:

$$\frac{R^2_{\boldsymbol{b},\hat{\mathbf{b}}_{\mathrm{SBLUP}}}}{R^2_{\mathbf{b}_k,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}}} = \frac{Nh^2+M}{Nh^2+M\left(1-R_k^2\right)} \qquad [40]$$

For a phenotype of SNP-heritability 0.5, with effective number of independent markers (independent genomic segments), M_{eff}, of ~60,000 and sample size, N, of 500,000, $R^2_{\boldsymbol{b},\hat{\mathbf{b}}_{\mathrm{SBLUP}}}$ from summary statistics in an independent reference sample will be 91% of the value of $R^2_{\mathbf{b}_k,\hat{\mathbf{b}}_{\mathrm{BLUP}_k}}$ if individual-level data were available. Likewise for a two-trait example where both traits have $h^2 = 0.5$ and N = 500,000, the accuracy of the multi-trait SBLUP predictor will also be 91% of the accuracy of the multi-trait BLUP predictor.

It should be noted that here we assume $\boldsymbol{L}$ to be a diagonal matrix, which will lead to a conservative estimate of the accuracy of SBLUP relative to the accuracy of BLUP, and that this estimate is in fact equivalent to the expected accuracy of a polygenic risk predictor based on marginal OLS effects[4]. In practice, approximating $\boldsymbol{L}$ through an external reference data set leads to SBLUP predictors which are more accurate than predictors based on marginal OLS effects, but less accurate than predictors based on BLUP effects.

Simulation study

To compare the accuracy of single-trait and multi-trait genetic predictors created from SNP effects obtained from both individual-level and summary statistic data, we conducted a simulation study based on real genotypes from the Kaiser Permanente study (Genetic Epidemiology Research on Adult Health and Aging: GERA cohort) and simulated phenotypes.

From the GERA cohort, we selected 50,000 individuals of European ancestry (for definitions of European individuals and quality control of the genotypic data see [215]). SNP genotype data was imputed to a 1000 genomes reference panel, using quality control (QC) procedures on the initial datasets of per-SNP missing data rate of < 0.01, minor allele frequency > 0.01, per-person missing data rate < 0.01, and Hardy-Weinberg disequilibrium p-value < $1x10^{-6}$. Imputation was performed in two stages. First, the target data was haplotyped using HAPI-

UR. Second, Impute2 was used to impute the haplotypes to the 1000 genomes reference panel (release 1, version 3). We then extracted best-guess genotypes at common SNPs typed in the HapMap 3 European sample with imputation info score > 0.5, missing data rate of < 0.01, minor allele frequency > 0.01, per-person missing data rate < 0.01, and Hardy-Weinberg disequilibrium p-value < $1x10^{-6}$. We conducted principal component analysis and removed individuals with principal eigenvector values that were > 7 SD from the mean. Finally, we removed one of any pair of individuals with estimated relatedness in a genetic relatedness matrix greater than a threshold of 0.05.

The Atherosclerosis Risk in Communities study (ARIC data) was used as an independent LD reference when estimating SBLUP SNP effects of Eq. [11]. 8744 European individuals were selected and the data was imputed and QC conducted in the same way as described above for the GERA cohort. We then reduced the SNPs used in both the GERA and ARIC cohorts to overlapping HapMap3 SNPs, which gave 557,034 SNPs that were used in the simulation study.

We then randomly assigned 20,000, 20,000 and 10,000 individuals from the GERA cohort to create three datasets: training set one, training set two, and a testing set. We simulated two genetically correlated traits by randomly selecting 2000 causal SNPs Effect sizes for the causal markers were simulated from a bivariate normal distribution with mean 0, variances of $\frac{h_1^2}{M}$ and $\frac{h_2^2}{M}$ and covariance of $r_G\sqrt{h_1^2 h_2^2}$. These effect sizes were then multiplied with the standardized genotype dosages (mean 0 and variance 1) to create a genetic value for each individual. Normally distributed environmental effects $e \sim N(0,\ 1-h^2)$ were added to this genetic value for each individual to create phenotypes with mean 0 and variance of 1. To remove any effects of population stratification, the simulated phenotypes were then regressed against the first 20 genetic principal components, and the residuals from this regression were used in all subsequent analyses.

In training set 1, we simulated trait 1 and we then estimated: (i) OLS SNP effects using Eq. [5] $\left(\hat{\mathbf{b}}_{\mathrm{OLS}}\right)$, (ii) BLUP SNP effects from the individual-level data using Eq. [8] $\left(\hat{\mathbf{b}}_{\mathrm{BLUP}}\right)$, and (iii) approximate SBLUP effects using the OLS SNP effects from Eq. [5] and the ARIC data as a reference $\left(\hat{\mathbf{b}}_{\mathrm{SBLUP}}\right)$. In training set 2, we simulated trait 2 and estimated $\hat{\mathbf{b}}_{\mathrm{OLS}}$, $\hat{\mathbf{b}}_{\mathrm{BLUP}}$, and $\hat{\mathbf{b}}_{\mathrm{SBLUP}}$ in the same manner. We then estimated multi-trait BLUP SNP effects using Eq.

[9] $\left(\hat{\mathbf{b}}_{\mathrm{MT-BLUP}}\right)$ from individual-level data by combining trait 1 from training set 1 and trait 2 from training set 2.

In the testing set, we then used the estimated SNP effects from the training sets to produce genetic predictors for both traits. Single trait genetic predictors were created for both simulated traits from (i) the OLS SNP effects $(\hat{\mathbf{g}}_{\mathrm{OLS}})$, (ii) the BLUP SNP effects $(\hat{\mathbf{g}}_{\mathrm{BLUP}})$, and (iii) the SBLUP SNP effects $(\hat{\mathbf{g}}_{\mathrm{SBLUP}})$. We then created multi-trait predictors where trait 1 was the focal trait from: (i) individual-level multi-trait BLUP predictor $(\hat{\mathbf{g}}_{\mathrm{MT-BLUP}})$, (ii) weighted multi-trait SBLUP predictor $(\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}})$**,** (iii) a weighted multi-trait BLUP predictor based individual-level single-trait BLUP estimates $(\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}})$, and (iv) a weighted multi-trait GWAS predictor based on GWAS OLS estimates $(\hat{\mathbf{g}}_{\mathrm{wMT-OLS}})$. We simulated phenotypic values for both traits using the same effect sizes as those used to generate the phenotypes in the training sets, and normally distributed environmental effects sampled independently for each trait as $e \sim N(0,\ 1-h^2)$.

We created two simulation scenarios. Heritability of the first and second trait, and genetic correlations were h_1^2 = 0.2, h_2^2 = 0.8, and r_G = 0.8, respectively in the first scenario and were h_1^2 = 0.5, h_2^2 = 0.5, and r_G = 0.5 respectively in the second scenario. In each setup 6 replicates were conducted, each with a different set of randomly selected causal markers. We then repeated all analyses on a permuted data set, where the values of the genotype matrix were permuted across all individuals, for each SNP. This creates a genotype matrix where the allele frequency distribution remains the same, but all LD structure is removed, allowing us to determine the degree to which differences between the simulations results are driven by the LD structure in the real genotype data. Finally, because prediction accuracy is expected to be reduced by the error introduced by using an external LD reference data set and a restricted LD window when implementing Eq. [8] (see above), we examined how changing the LD reference, and restricting the LD window size, influences to optimal value of shrinkage parameter λ, when implementing Eq. [8], (see **Figure 33**).

Application to PGC schizophrenia and bipolar disorder

We then applied our approach to the schizophrenia (SCZ) and bipolar disorder (BIP) samples from both wave 1 and wave 2 data of the Psychiatric Genomics Consortium (PGC1

and PGC2). A description of the data collection and imputation of the SNP genotype data can be found elsewhere [10,58,216].

We selected these two disorders because there is a high genetic correlation between them (estimate for r_G between schizophrenia and bipolar disorder using ldsc: 0.72, SE: 0.03; estimated using meta-analysis of all PGC2 schizophrenia and bipolar cohorts, excluding cohorts which were used as test set in the initial PGC1 analysis) and it enabled us to draw a direct comparison between the approach described here and a previous study which estimated multi-trait BLUP SNP effects $(\hat{\mathbf{b}}_{\mathrm{MT-BLUP}})$ from individual-level data in an approach equivalent to Eq. [9]. The previous study used PGC1 data in the training set and selected 4 cohorts for schizophrenia and 3 cohorts for bipolar disorder as test sets. For schizophrenia, the training set comprised 17 cohorts (8826 cases, 6106 controls) and for bipolar disorder the training set comprised 11 cohorts (5867 cases, 3328 controls). The test set of 4 cohorts for schizophrenia contained 4068 cases and 5471 controls, and the test set of 3 cohorts for bipolar disorder contained 2029 cases and 5338 controls. The analyses on the PGC1 data were performed on 745,705 HapMap3 SNPs in common across all datasets. To have a direct comparison to our previous study, we began by re-analysing the same PGC1 training set data to estimate: (i) OLS SNP effects using Eq. [5] $(\hat{\mathbf{b}}_{\mathrm{OLS}})$, (ii) BLUP SNP effects from the individual-level data using Eq. [8] $(\hat{\mathbf{b}}_{\mathrm{BLUP}})$, and (iii) approximate SBLUP effects using the OLS SNP effects from Eq. [5] and the ARIC data as a reference $(\hat{\mathbf{b}}_{\mathrm{SBLUP}})$ using Eq. [11]. For the estimation of schizophrenia SBLUP effects, λ was set to 1,100,000, corresponding roughly to 1,000,000 markers and an observed scale SNP-heritability estimate of 0.47 and for the estimation of bipolar disorder SBLUP effects, lambda was set to 1,200,000, corresponding roughly to 1,000,000 markers and an observed scale SNP-heritability estimate of 0.45. For the four SCZ testing cohorts and the three BIP testing cohorts used in the previous study, we created: (i) weighted multi-trait SBLUP predictors $(\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}})$, (ii) weighted multi-trait BLUP predictor based individual-level single-trait BLUP estimates $(\hat{\mathbf{g}}_{\mathrm{wMT-BLUP}})$, and (iii) weighted multi-trait GWAS predictor based on GWAS OLS estimates $(\hat{\mathbf{g}}_{\mathrm{wMT-OLS}})$. We then compared the prediction accuracy we obtained using the weighted multi-trait SBLUP predictors to the individual-level multi-trait BLUP predictor $(\hat{\mathbf{g}}_{\mathrm{MT-BLUP}})$ used in the previous study [66].

We then extended our analysis to the PGC2 dataset. There were 36 cohorts for schizophrenia (26412 cases and 32440 controls in total) and 23 cohorts for bipolar disorder

(18865 cases and 30460 controls in total) available to us. The number of SNPs used in the PGC2 analyses varied between cohorts. Summary statistics for each of the PGC2 cohorts was available to an imputed SNP set of more than 10,000,000 SNPs. After intersecting this set of SNPs with the HapMap3 SNPs and the ARIC SNPs, 932,344 SNPs remained which were used to create predictors.

We applied a cross-validation approach as we observed that prediction accuracy as well as accuracy differences between predictors can be highly dependent on the choice of the test set in the extended PGC2 dataset (**Figure 35** and **Figure 36**), which is supported by previous results showing highly variable prediction accuracy across cohorts in the PGC2 dataset [58]. A cross-validation approach allowed us to get a more robust estimate of the increase of prediction accuracy achieved by our multi-trait prediction method compared to a single trait predictor. We employed a leave-1-out-cross-validation approach, where for each test set cohort, all cohorts of the same disease without any highly related individuals were chosen to be in the training set for the single-trait predictor, and all cohorts of both diseases without any highly related individuals were chosen to be in the training set for the multi-trait predictor. To identify pairs of cohorts with highly related individuals, genetic relatedness for all pairs of individuals (across all pairs of cohorts) was calculated based on chromosome 22, and whenever at least one pair of individuals had relatedness greater than 0.8, that pair of cohorts was not simultaneously used in the training set and the test set.

The full genotypes from the PGC2 cohorts that were used as test sets underwent stringent quality control and only comprised 458,744 to 860,576 SNPs for schizophrenia and 556,278 to 859,034 SNPs for bipolar disorder. We refrained from using the intersection between all these cohorts to not reduce the number of SNPs used in prediction by too much. This meant that different iterations in the cross-validations were based on predictions using a different number of SNPs. However, each comparison between a single trait predictor and a multi-trait predictor is based on the same number of SNPs.

In each iteration of the cross-validation, a different cohort acts as the test set and a different set of cohorts comprises the training set. To create a predictor from a particular set of cohorts, we first had to obtain effect size estimates from this particular set of cohorts. This is achieved by performing a meta-analysis of the summary statistics of the cohorts that comprise the training set. The meta-analysed beta values b_{META} are calculated as:

$$b_{META} = \frac{\sum_S \frac{b_S}{SE_S^2}}{\sum_S \frac{1}{SE_S^2}} \quad [41]$$

where b_s is the effect size in cohort *s* and SE_s is the standard error in cohort *s*. Conversion between beta values and odds ratios (OR) simply follows the equality $b = \log(OR)$. The weights derived for each trait make assumptions about the variance of SNP effects. We found that in the summary statistics we used, the observed variance across SNP effects often departed from the expected value. To correct for that, we scaled the SNP effect estimates for each trait to have a variance of one and multiplied the weights for the unscaled SNP effects by the expected standard deviation across all SNPs.

We created approximate SBLUP effects $(\hat{\mathbf{b}}_{\mathrm{SBLUP}})$ using the OLS SNP effects from Eq. [5] and the ARIC data as an LD reference using Eq. [11] and set the shrinkage parameter, λ, to 1,300,000 for schizophrenia and to 2,000,000 for bipolar disorder, corresponding to observed scale SNP-heritability estimates of 0.43 and 0.33 for schizophrenia and bipolar disorder, respectively. We then used the PLINK "--score" function to turn SNP effects $(\hat{\mathbf{b}}_{\mathrm{SBLUP}}, \hat{\mathbf{b}}_{\mathrm{GWAS}})$ into individual predictors $(\hat{\mathbf{g}}_{\mathrm{SBLUP}}, \hat{\mathbf{g}}_{\mathrm{GWAS}})$ for each meta-analysed schizophrenia or bipolar disorder cross-validation set. For the multi-trait weighting, we estimated the heritability of schizophrenia and bipolar disorder and their genetic correlation using LD score regression from publicly available PGC2 schizophrenia summary statistics and the PGC1 bipolar disorder summary statistics. These estimates were then used to calculate the index weights of Eq. [15] for the weighted multi-trait SBLUP predictors $(\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}}, \hat{\mathbf{g}}_{\mathrm{wMT-GWAS}})$ of SCZ and BIP, and these were not altered between different cross-validation sets.

To test the degree to which the choice of weights affects the accuracy of the multi-trait predictor, we compared the accuracy of multi-trait predictors based on a spectrum of other weights (**Figure 35** and **Figure 36**). For this, we took advantage of two things: First, when individual predictors $(\hat{\mathbf{g}}_{\mathrm{SBLUP}}, \hat{\mathbf{g}}_{\mathrm{GWAS}})$ are weighted rather than SNP effects $(\hat{\mathbf{b}}_{\mathrm{SBLUP}}, \hat{\mathbf{b}}_{\mathrm{GWAS}})$, the conversion from SNP effects to individual effects does not have to be repeated for different weights. Second, the scaling of a predictor does not influence its accuracy in terms of correlation between prediction and outcome. Therefore, rather than testing each combination of weights of schizophrenia and bipolar disorder, it is sufficient to vary the relative weight of schizophrenia to bipolar disorder to explore the whole range of possible

multi-trait predictors for these two traits. For each test cohort, this enabled us to test whether the weights of our multi-trait predictor derived from theory deviate from the weights that would result in the highest prediction accuracy for that dataset.

Application to wide range of phenotypes in the UK Biobank study

We applied our approach to a large range of phenotypes for which GWAS summary statistics are publicly available. We started with GWAS summary statistics for 46 phenotypes. However, in some circumstance the same studies (i.e., based on the same individuals) had generated summary statistics for multiple similar phenotypes, so we chose only one phenotype per study, which left us with 34 phenotypes. For example, out of "Cigarettes per day" and "Smoking Ever" we only selected the latter to have only one trait for smoking. We used 112,338 unrelated individuals of European descent in the UK biobank data as the testing set. We paired 6 phenotypes out of the 34 summary statistic phenotypes to phenotypes in the UK Biobank: Height, BMI, fluid intelligence score, depression, angina and diabetes. The first three are quantitative traits and the latter three are disease traits for which we could identify at least 1000 cases in the UK Biobank data. For details see **Table 15**.

For the disease traits, we used the self-reported diagnoses rather than ICD10 diagnoses, as they tend to have larger sample sizes. For depression, we used a more refined definition of cases and controls, where individuals were not counted as cases if they had any history of psychiatric symptoms or diagnoses other than depression, or if they were prescribed drugs that are indicative of such diagnoses. Individuals were selected as controls only when there was an absence of any psychiatric symptoms or diagnoses, and only when they were not prescribed any drugs that could be indicative of such diagnoses. All 6 traits in the UK Biobank were corrected for age, sex and the first 10 principal components by regressing the phenotype on these covariates and using the residuals from that regression for further analysis. For each trait, the SNPs that went into the analysis were based on the overlap between the GWAS summary statistics, the HapMap3 SNPs, the GERA data set, which was used as an LD reference in the SBLUP analysis, and the imputed SNPs from the UK Biobank. (For details on the QC process and imputation, see URLs). Depending on the trait, the total number of SNPs ranged from around 660,000 to around 930,000.

We created single-trait ($\hat{\mathbf{g}}_{\mathrm{SBLUP}}$) as well as multi-trait ($\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}}$) predictors for the 6 paired phenotypes. To create SBLUP SNP effects ($\hat{\mathbf{b}}_{\mathrm{SBLUP}}$) from summary statistic trait we used a λ value of $M(1-h^2_{SNP_k})/h^2_{SNP_k}$ for each trait *k*, where *M* is assumed to be 1,000,000. As LD reference set we used a random subset of 10,000 people of European descent from the GERA dataset, and we set the LD window size to 2,000 kb. We then used the PLINK "--score" function to turn SNP effects ($\hat{\mathbf{b}}_{\mathrm{SBLUP}}$) into individual predictors ($\hat{\mathbf{g}}_{\mathrm{SBLUP}}$) for each trait. For the multi-trait weighting, we used LD score regression to calculate SNP-heritability and genetic correlation between all pairs of cohorts. For dichotomous disease traits SNP-heritability was calculated on the observed scale. For each phenotype for which a multi-trait predictor was created, we selected all phenotypes which had a genetic correlation estimate significantly different from 0 at p = 0.05 with the focal trait, as well as the focal trait itself. The summary statistics based single-trait SBLUP predictors of the selected phenotypes were then combined into multi-trait SBLUP ($\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}}$) predictors. The weights for each phenotype were calculated according to Equation [15]. These weights require the single-trait predictors to have exactly the right variance. Since the summary statistics data slightly diverged from this expectation, we scaled each single-trait SBLUP predictor to have mean 0 and variance 1 and then multiplied it with its expected standard deviation, to ensure everything is on exactly the correct scale.

We compared the performance of the multi-trait predictors ($\hat{\mathbf{g}}_{\mathrm{wMT-SBLUP}}$) not only to the performance of the single-trait predictor ($\hat{\mathbf{g}}_{\mathrm{SBLUP}}$) for the same trait, but also to the performance of all other (cross-trait) single-trait predictors for the traits that exhibited significant r_G with the focal trait (**Figure 39**). This is appropriate because in some traits the single-trait predictor from the same trait is not the most accurate single-trait predictor.

Code availability

Code is available from https://github.com/uqrmaie1/smtpred.

Data availability

PGC summary statistics data is available from http://www.med.unc.edu/pgc/results-and-downloads.
For UK Biobank data, see https://www.ukbiobank.ac.uk/.

URLs

GCTA, http://cnsgenomics.com/software/gcta/
LDSC, https://github.com/bulik/ldsc
MTG2, https://sites.google.com/site/honglee0707/mtg/
LDpred, https://github.com/bvilhjal/ldpred/
UK Biobank, http://www.ukbiobank.ac.uk/
PLINK2, http://www.cog-genomics.org/plink2

Acknowledgements

The University of Queensland group is supported by the Australian Research Council (Discovery Project 160103860), the Australian National Health and Medical Research Council (NHMRC grants 1087889, 1080157, 1048853, 1050218, 1078901 and 1078037), and the National Institute of Health (NIH grants R21ESO25052-01 and PO1GMO99568). J.Y. is supported by a Charles and Sylvia Viertel Senior Medical Research Fellowship. We thank the all the participants and researchers of the many cohort studies that make this work possible, as well as our colleagues within The University of Queensland's Program for Complex Trait Genomics and the Queensland Brain Institute IT team for comments and suggestions and technical support. The UK Biobank research was conducted using the UK Biobank Resource under project 12514. Statistical analyses of PGC data were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) hosted by SURFsara and financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

Numerous (>100) grants from government agencies along with substantial private and foundation support worldwide enabled the collection of phenotype and genotype data, without which this research would not be possible; grant numbers are listed in primary PGC publications.

Corresponding author

Correspondence to: Robert M Maier r.maier@uq.edu.au, Matthew R Robinson matthew.robinson@unil.ch, or Peter M Visscher peter.visscher@uq.edu.au

Table 17: LDSC r_G estimates

Shaded lines highlight pairs with p < 0.05

Trait 1	Trait 2	rg	SE	p
ADHD	Agreeableness	-1.02	1.09	0.35
ADHD	Alzheimers	-0.15	0.24	0.52
ADHD	Autism	-0.22	0.16	0.17
ADHD	Bipolar	0.21	0.14	0.13
ADHD	Birth Length	-0.19	0.25	0.45
ADHD	Birth Weight	0.10	0.21	0.65
ADHD	BMI	0.31	0.12	0.01
ADHD	CAD	0.33	0.12	0.00
ADHD	Childhood Obesity	0.00	0.13	1.00
ADHD	Conscientiousness	-0.44	0.40	0.27
ADHD	Crohns	0.12	0.11	0.28
ADHD	Depression	0.06	0.19	0.76
ADHD	Diabetes	0.20	0.17	0.25
ADHD	Education	-0.31	0.14	0.02
ADHD	Extraversion	-0.43	0.67	0.51
ADHD	Glucose	-0.06	0.15	0.69
ADHD	Head Circumference	-0.59	0.31	0.06
ADHD	Height	-0.02	0.08	0.81
ADHD	Inflammatory Bowel Disease	0.17	0.11	0.14
ADHD	Insulin	-0.05	0.19	0.77
ADHD	IQ	-0.12	0.23	0.61
ADHD	MND	0.12	0.30	0.68
ADHD	Neuroticism	0.01	0.21	0.97
ADHD	Openness	-0.44	0.38	0.25
ADHD	Osteoporosis	0.23	0.14	0.11
ADHD	Pubertal Growth	-0.07	0.15	0.61
ADHD	Rheumatoid Arthritis	0.02	0.12	0.88

ADHD	Schizophrenia	0.11	0.09	0.23
ADHD	Smoking	0.27	0.15	0.06
ADHD	Tanner	0.56	0.39	0.16
ADHD	Triglycerides	0.10	0.09	0.27
ADHD	Ulcerative Colitis	0.22	0.13	0.10
ADHD	Waist Hip Ratio	0.32	0.13	0.02
Agreeableness	Alzheimers	0.82	0.74	0.26
Agreeableness	Autism	-0.09	0.31	0.76
Agreeableness	Bipolar	-0.14	0.33	0.67
Agreeableness	Birth Length	0.10	0.36	0.77
Agreeableness	Birth Weight	-0.31	0.47	0.51
Agreeableness	BMI	-0.27	0.23	0.23
Agreeableness	CAD	-0.46	0.38	0.23
Agreeableness	Childhood Obesity	0.42	0.43	0.32
Agreeableness	Conscientiousness	0.53	0.58	0.36
Agreeableness	Crohns	-0.45	0.39	0.25
Agreeableness	Depression	-0.74	0.75	0.32
Agreeableness	Diabetes	-0.48	0.45	0.28
Agreeableness	Education	0.58	0.38	0.12
Agreeableness	Extraversion	1.19	1.15	0.30
Agreeableness	Glucose	-0.14	0.29	0.64
Agreeableness	Head Circumference	0.16	0.47	0.73
Agreeableness	Height	0.00	0.14	0.97
Agreeableness	Inflammatory Bowel Disease	-0.05	0.24	0.84
Agreeableness	Insulin	-0.33	0.45	0.47
Agreeableness	IQ	0.54	0.84	0.52
Agreeableness	MND	-0.45	0.55	0.42
Agreeableness	Neuroticism	-1.38	1.18	0.24
Agreeableness	Openness	1.15	1.01	0.25
Agreeableness	Osteoporosis	-0.19	0.39	0.62

Agreeableness	Pubertal Growth	-0.14	0.32	0.67
Agreeableness	Rheumatoid Arthritis	-0.40	0.40	0.32
Agreeableness	Schizophrenia	-0.41	0.35	0.25
Agreeableness	Smoking	-0.18	0.32	0.56
Agreeableness	Tanner	-0.60	0.59	0.31
Agreeableness	Triglycerides	-0.02	0.14	0.91
Agreeableness	Ulcerative Colitis	0.25	0.36	0.49
Agreeableness	Waist Hip Ratio	-0.25	0.24	0.31
Alzheimers	Autism	-0.01	0.12	0.91
Alzheimers	Bipolar	0.05	0.10	0.64
Alzheimers	Birth Length	-0.11	0.13	0.37
Alzheimers	Birth Weight	-0.05	0.13	0.70
Alzheimers	BMI	-0.02	0.06	0.70
Alzheimers	CAD	-0.03	0.08	0.69
Alzheimers	Childhood Obesity	-0.12	0.10	0.24
Alzheimers	Conscientiousness	0.02	0.17	0.92
Alzheimers	Crohns	-0.09	0.09	0.33
Alzheimers	Depression	0.18	0.16	0.25
Alzheimers	Diabetes	0.01	0.12	0.94
Alzheimers	Education	-0.37	0.11	0.00
Alzheimers	Extraversion	-0.07	0.31	0.82
Alzheimers	Glucose	0.21	0.11	0.05
Alzheimers	Head Circumference	-0.27	0.17	0.12
Alzheimers	Height	-0.11	0.05	0.04
Alzheimers	Inflammatory Bowel Disease	-0.11	0.09	0.23
Alzheimers	Insulin	-0.20	0.13	0.12
Alzheimers	IQ	-0.36	0.12	0.00
Alzheimers	MND	0.16	0.21	0.45
Alzheimers	Neuroticism	0.11	0.17	0.50
Alzheimers	Openness	-0.08	0.16	0.63

Alzheimers	Osteoporosis	-0.18	0.10	0.07
Alzheimers	Pubertal Growth	-0.32	0.12	0.01
Alzheimers	Rheumatoid Arthritis	-0.05	0.11	0.64
Alzheimers	Schizophrenia	0.06	0.06	0.35
Alzheimers	Smoking	0.01	0.10	0.90
Alzheimers	Tanner	-0.02	0.19	0.90
Alzheimers	Triglycerides	-0.04	0.06	0.58
Alzheimers	Ulcerative Colitis	-0.15	0.11	0.15
Alzheimers	Waist Hip Ratio	0.01	0.07	0.86
Autism	Bipolar	0.05	0.08	0.55
Autism	Birth Length	0.19	0.11	0.09
Autism	Birth Weight	0.08	0.10	0.42
Autism	BMI	0.01	0.04	0.82
Autism	CAD	-0.05	0.05	0.32
Autism	Childhood Obesity	-0.06	0.08	0.44
Autism	Conscientiousness	-0.24	0.17	0.16
Autism	Crohns	-0.04	0.06	0.47
Autism	Depression	0.15	0.11	0.20
Autism	Diabetes	-0.01	0.09	0.96
Autism	Education	0.31	0.05	0.00
Autism	Extraversion	-0.39	0.33	0.24
Autism	Glucose	-0.17	0.08	0.02
Autism	Head Circumference	0.07	0.12	0.53
Autism	Height	-0.07	0.04	0.06
Autism	Inflammatory Bowel Disease	-0.02	0.05	0.64
Autism	Insulin	0.08	0.10	0.42
Autism	IQ	0.43	0.12	0.00
Autism	MND	-0.16	0.18	0.35
Autism	Neuroticism	0.02	0.13	0.85
Autism	Openness	0.48	0.16	0.00

Autism	Osteoporosis	0.03	0.07	0.65
Autism	Pubertal Growth	-0.08	0.08	0.30
Autism	Rheumatoid Arthritis	-0.10	0.09	0.23
Autism	Schizophrenia	0.20	0.05	0.00
Autism	Smoking	0.08	0.08	0.31
Autism	Tanner	0.09	0.17	0.58
Autism	Triglycerides	0.05	0.05	0.28
Autism	Ulcerative Colitis	0.01	0.07	0.92
Autism	Waist Hip Ratio	0.08	0.06	0.15
Bipolar	Birth Length	0.00	0.08	0.97
Bipolar	Birth Weight	0.01	0.09	0.94
Bipolar	BMI	-0.02	0.04	0.49
Bipolar	CAD	0.07	0.04	0.09
Bipolar	Childhood Obesity	-0.04	0.06	0.49
Bipolar	Conscientiousness	-0.33	0.15	0.03
Bipolar	Crohns	0.20	0.05	0.00
Bipolar	Depression	0.53	0.09	0.00
Bipolar	Diabetes	0.04	0.06	0.53
Bipolar	Education	0.23	0.04	0.00
Bipolar	Extraversion	-0.07	0.23	0.77
Bipolar	Glucose	0.04	0.07	0.53
Bipolar	Head Circumference	0.06	0.10	0.58
Bipolar	Height	-0.01	0.03	0.65
Bipolar	Inflammatory Bowel Disease	0.18	0.05	0.00
Bipolar	Insulin	0.00	0.08	0.99
Bipolar	IQ	0.07	0.08	0.37
Bipolar	MND	0.13	0.11	0.26
Bipolar	Neuroticism	0.08	0.10	0.40
Bipolar	Openness	0.08	0.11	0.48
Bipolar	Osteoporosis	-0.05	0.06	0.42

Bipolar	Pubertal Growth	-0.08	0.08	0.29
Bipolar	Rheumatoid Arthritis	-0.06	0.07	0.40
Bipolar	Schizophrenia	0.81	0.04	0.00
Bipolar	Smoking	0.07	0.06	0.25
Bipolar	Tanner	0.18	0.13	0.18
Bipolar	Triglycerides	0.00	0.04	0.93
Bipolar	Ulcerative Colitis	0.15	0.06	0.01
Bipolar	Waist Hip Ratio	0.00	0.05	0.98
Birth Length	Birth Weight	0.69	0.08	0.00
Birth Length	BMI	0.04	0.04	0.35
Birth Length	CAD	-0.21	0.06	0.00
Birth Length	Childhood Obesity	0.03	0.09	0.71
Birth Length	Conscientiousness	-0.03	0.17	0.86
Birth Length	Crohns	0.00	0.07	0.96
Birth Length	Depression	-0.10	0.12	0.39
Birth Length	Diabetes	-0.24	0.10	0.02
Birth Length	Education	0.14	0.06	0.02
Birth Length	Extraversion	0.22	0.24	0.37
Birth Length	Glucose	-0.12	0.08	0.14
Birth Length	Head Circumference	0.55	0.13	0.00
Birth Length	Height	0.51	0.05	0.00
Birth Length	Inflammatory Bowel Disease	0.10	0.07	0.16
Birth Length	Insulin	0.04	0.11	0.71
Birth Length	IQ	-0.01	0.10	0.93
Birth Length	MND	0.02	0.14	0.87
Birth Length	Neuroticism	0.06	0.14	0.68
Birth Length	Openness	-0.10	0.14	0.49
Birth Length	Osteoporosis	0.02	0.08	0.76
Birth Length	Pubertal Growth	0.49	0.08	0.00
Birth Length	Rheumatoid Arthritis	-0.09	0.08	0.29

Birth Length	Schizophrenia	0.01	0.05	0.92
Birth Length	Smoking	-0.10	0.09	0.26
Birth Length	Tanner	-0.37	0.17	0.03
Birth Length	Triglycerides	-0.06	0.05	0.18
Birth Length	Ulcerative Colitis	0.19	0.09	0.03
Birth Length	Waist Hip Ratio	-0.06	0.06	0.25
Birth Weight	BMI	0.11	0.04	0.01
Birth Weight	CAD	-0.14	0.06	0.02
Birth Weight	Childhood Obesity	0.11	0.08	0.16
Birth Weight	Conscientiousness	-0.10	0.16	0.53
Birth Weight	Crohns	0.09	0.08	0.27
Birth Weight	Depression	-0.04	0.12	0.75
Birth Weight	Diabetes	-0.35	0.10	0.00
Birth Weight	Education	0.11	0.05	0.04
Birth Weight	Extraversion	-0.13	0.34	0.71
Birth Weight	Glucose	-0.20	0.09	0.03
Birth Weight	Head Circumference	0.42	0.14	0.00
Birth Weight	Height	0.43	0.05	0.00
Birth Weight	Inflammatory Bowel Disease	0.05	0.08	0.51
Birth Weight	Insulin	-0.19	0.13	0.13
Birth Weight	IQ	0.13	0.13	0.31
Birth Weight	MND	0.25	0.16	0.12
Birth Weight	Neuroticism	-0.02	0.14	0.90
Birth Weight	Openness	0.07	0.15	0.63
Birth Weight	Osteoporosis	0.09	0.08	0.25
Birth Weight	Pubertal Growth	0.29	0.09	0.00
Birth Weight	Rheumatoid Arthritis	-0.09	0.09	0.33
Birth Weight	Schizophrenia	0.03	0.06	0.56
Birth Weight	Smoking	-0.06	0.09	0.53
Birth Weight	Tanner	0.08	0.20	0.68

Birth Weight	Triglycerides	-0.11	0.05	0.03
Birth Weight	Ulcerative Colitis	-0.01	0.09	0.93
Birth Weight	Waist Hip Ratio	-0.19	0.07	0.01
BMI	CAD	0.21	0.03	0.00
BMI	Childhood Obesity	0.77	0.04	0.00
BMI	Conscientiousness	0.04	0.07	0.55
BMI	Crohns	0.02	0.03	0.55
BMI	Depression	-0.07	0.05	0.19
BMI	Diabetes	0.34	0.05	0.00
BMI	Education	-0.28	0.02	0.00
BMI	Extraversion	0.17	0.14	0.23
BMI	Glucose	0.26	0.05	0.00
BMI	Head Circumference	0.17	0.05	0.00
BMI	Height	-0.06	0.02	0.00
BMI	Inflammatory Bowel Disease	-0.04	0.03	0.21
BMI	Insulin	0.64	0.06	0.00
BMI	IQ	-0.16	0.05	0.00
BMI	MND	0.05	0.07	0.45
BMI	Neuroticism	0.04	0.06	0.54
BMI	Openness	0.04	0.06	0.51
BMI	Osteoporosis	-0.06	0.03	0.07
BMI	Pubertal Growth	0.20	0.04	0.00
BMI	Rheumatoid Arthritis	0.00	0.03	0.95
BMI	Schizophrenia	-0.09	0.02	0.00
BMI	Smoking	0.20	0.03	0.00
BMI	Tanner	0.33	0.10	0.00
BMI	Triglycerides	0.19	0.03	0.00
BMI	Ulcerative Colitis	-0.09	0.03	0.01
BMI	Waist Hip Ratio	-0.08	0.03	0.01
CAD	Childhood Obesity	0.18	0.05	0.00

CAD	Conscientiousness	-0.13	0.11	0.20
CAD	Crohns	0.08	0.04	0.05
CAD	Depression	0.23	0.07	0.00
CAD	Diabetes	0.36	0.06	0.00
CAD	Education	-0.30	0.03	0.00
CAD	Extraversion	-0.03	0.17	0.86
CAD	Glucose	0.12	0.05	0.02
CAD	Head Circumference	-0.06	0.08	0.43
CAD	Height	-0.11	0.02	0.00
CAD	Inflammatory Bowel Disease	0.07	0.04	0.10
CAD	Insulin	0.25	0.07	0.00
CAD	IQ	-0.07	0.07	0.36
CAD	MND	-0.09	0.10	0.39
CAD	Neuroticism	0.15	0.07	0.04
CAD	Openness	-0.04	0.08	0.59
CAD	Osteoporosis	-0.06	0.04	0.17
CAD	Pubertal Growth	-0.01	0.05	0.81
CAD	Rheumatoid Arthritis	0.04	0.04	0.30
CAD	Schizophrenia	0.00	0.03	0.96
CAD	Smoking	0.22	0.05	0.00
CAD	Tanner	0.39	0.13	0.00
CAD	Triglycerides	0.26	0.04	0.00
CAD	Ulcerative Colitis	0.03	0.05	0.61
CAD	Waist Hip Ratio	0.18	0.03	0.00
Childhood Obesity	Conscientiousness	0.18	0.15	0.22
Childhood Obesity	Crohns	-0.02	0.06	0.72
Childhood Obesity	Depression	-0.10	0.10	0.28
Childhood Obesity	Diabetes	0.24	0.07	0.00
Childhood Obesity	Education	-0.20	0.04	0.00
Childhood Obesity	Extraversion	0.11	0.21	0.60

Childhood Obesity	Glucose	0.06	0.07	0.38
Childhood Obesity	Head Circumference	0.31	0.10	0.00
Childhood Obesity	Height	-0.01	0.04	0.82
Childhood Obesity	Inflammatory Bowel Disease	-0.16	0.05	0.00
Childhood Obesity	Insulin	0.29	0.09	0.00
Childhood Obesity	IQ	-0.09	0.09	0.29
Childhood Obesity	MND	-0.03	0.15	0.86
Childhood Obesity	Neuroticism	0.08	0.11	0.46
Childhood Obesity	Openness	0.10	0.12	0.40
Childhood Obesity	Osteoporosis	-0.11	0.07	0.09
Childhood Obesity	Pubertal Growth	0.30	0.06	0.00
Childhood Obesity	Rheumatoid Arthritis	0.06	0.06	0.35
Childhood Obesity	Schizophrenia	-0.05	0.04	0.18
Childhood Obesity	Smoking	0.18	0.07	0.01
Childhood Obesity	Tanner	0.51	0.17	0.00
Childhood Obesity	Triglycerides	0.11	0.04	0.02
Childhood Obesity	Ulcerative Colitis	-0.23	0.06	0.00
Childhood Obesity	Waist Hip Ratio	-0.14	0.04	0.00
Conscientiousness	Crohns	0.24	0.12	0.05
Conscientiousness	Depression	-0.56	0.20	0.01
Conscientiousness	Diabetes	0.16	0.15	0.28
Conscientiousness	Education	0.06	0.09	0.50
Conscientiousness	Extraversion	0.53	0.41	0.20
Conscientiousness	Glucose	0.11	0.14	0.43
Conscientiousness	Head Circumference	-0.03	0.21	0.89
Conscientiousness	Height	-0.07	0.07	0.33
Conscientiousness	Inflammatory Bowel Disease	0.22	0.11	0.05
Conscientiousness	Insulin	0.04	0.17	0.80
Conscientiousness	IQ	0.14	0.17	0.43

Conscientiousness	MND	0.07	0.23	0.78
Conscientiousness	Neuroticism	-0.66	0.26	0.01
Conscientiousness	Openness	0.30	0.24	0.22
Conscientiousness	Osteoporosis	-0.04	0.13	0.74
Conscientiousness	Pubertal Growth	-0.17	0.16	0.30
Conscientiousness	Rheumatoid Arthritis	-0.23	0.15	0.11
Conscientiousness	Schizophrenia	-0.14	0.09	0.13
Conscientiousness	Smoking	-0.29	0.14	0.04
Conscientiousness	Tanner	-0.31	0.27	0.25
Conscientiousness	Triglycerides	-0.01	0.07	0.86
Conscientiousness	Ulcerative Colitis	0.16	0.14	0.23
Conscientiousness	Waist Hip Ratio	-0.13	0.09	0.17
Crohns	Depression	0.04	0.07	0.54
Crohns	Diabetes	0.04	0.06	0.42
Crohns	Education	-0.06	0.04	0.13
Crohns	Extraversion	0.32	0.24	0.17
Crohns	Glucose	-0.06	0.06	0.27
Crohns	Head Circumference	0.06	0.09	0.51
Crohns	Height	0.05	0.03	0.11
Crohns	Inflammatory Bowel Disease	0.95	0.02	0.00
Crohns	Insulin	-0.01	0.07	0.89
Crohns	IQ	-0.11	0.10	0.24
Crohns	MND	0.15	0.11	0.19
Crohns	Neuroticism	-0.01	0.07	0.92
Crohns	Openness	-0.10	0.11	0.35
Crohns	Osteoporosis	0.16	0.05	0.00
Crohns	Pubertal Growth	0.06	0.06	0.33
Crohns	Rheumatoid Arthritis	0.04	0.07	0.58
Crohns	Schizophrenia	0.11	0.03	0.00
Crohns	Smoking	-0.02	0.06	0.68

Crohns	Tanner	-0.04	0.11	0.70
Crohns	Triglycerides	0.05	0.04	0.22
Crohns	Ulcerative Colitis	0.68	0.06	0.00
Crohns	Waist Hip Ratio	0.00	0.04	0.96
Depression	Diabetes	0.06	0.10	0.51
Depression	Education	-0.09	0.07	0.18
Depression	Extraversion	-0.87	0.48	0.07
Depression	Glucose	-0.10	0.09	0.24
Depression	Head Circumference	0.02	0.14	0.89
Depression	Height	-0.10	0.04	0.01
Depression	Inflammatory Bowel Disease	0.13	0.08	0.08
Depression	Insulin	-0.10	0.12	0.39
Depression	IQ	-0.04	0.14	0.76
Depression	MND	0.19	0.20	0.34
Depression	Neuroticism	1.15	0.19	0.00
Depression	Openness	0.16	0.16	0.30
Depression	Osteoporosis	0.15	0.09	0.09
Depression	Pubertal Growth	-0.36	0.11	0.00
Depression	Rheumatoid Arthritis	0.04	0.08	0.65
Depression	Schizophrenia	0.51	0.06	0.00
Depression	Smoking	0.23	0.10	0.02
Depression	Tanner	-0.05	0.20	0.80
Depression	Triglycerides	0.09	0.05	0.09
Depression	Ulcerative Colitis	0.21	0.10	0.03
Depression	Waist Hip Ratio	0.16	0.06	0.01
Diabetes	Education	-0.17	0.05	0.00
Diabetes	Extraversion	-0.21	0.22	0.33
Diabetes	Glucose	0.60	0.09	0.00
Diabetes	Head Circumference	-0.09	0.11	0.42
Diabetes	Height	-0.01	0.04	0.82

Diabetes	Inflammatory Bowel Disease	0.04	0.05	0.42
Diabetes	Insulin	0.50	0.11	0.00
Diabetes	IQ	-0.16	0.11	0.12
Diabetes	MND	0.07	0.13	0.61
Diabetes	Neuroticism	0.10	0.12	0.41
Diabetes	Openness	-0.14	0.11	0.21
Diabetes	Osteoporosis	-0.20	0.07	0.01
Diabetes	Pubertal Growth	0.12	0.07	0.10
Diabetes	Rheumatoid Arthritis	-0.04	0.07	0.61
Diabetes	Schizophrenia	0.00	0.04	0.98
Diabetes	Smoking	-0.01	0.07	0.93
Diabetes	Tanner	0.29	0.14	0.04
Diabetes	Triglycerides	0.30	0.05	0.00
Diabetes	Ulcerative Colitis	0.05	0.07	0.48
Diabetes	Waist Hip Ratio	0.26	0.05	0.00
Education	Extraversion	0.24	0.15	0.10
Education	Glucose	-0.18	0.05	0.00
Education	Head Circumference	0.28	0.07	0.00
Education	Height	0.14	0.02	0.00
Education	Inflammatory Bowel Disease	-0.05	0.04	0.24
Education	Insulin	-0.36	0.06	0.00
Education	IQ	0.70	0.07	0.00
Education	MND	-0.05	0.08	0.54
Education	Neuroticism	-0.46	0.08	0.00
Education	Openness	0.52	0.09	0.00
Education	Osteoporosis	-0.02	0.04	0.60
Education	Pubertal Growth	0.13	0.04	0.00
Education	Rheumatoid Arthritis	-0.16	0.05	0.00
Education	Schizophrenia	0.09	0.03	0.00

Education	Smoking	-0.33	0.05	0.00
Education	Tanner	-0.13	0.09	0.16
Education	Triglycerides	-0.18	0.03	0.00
Education	Ulcerative Colitis	-0.01	0.05	0.91
Education	Waist Hip Ratio	-0.21	0.03	0.00
Extraversion	Glucose	-0.10	0.27	0.72
Extraversion	Head Circumference	-0.52	0.43	0.23
Extraversion	Height	0.05	0.11	0.65
Extraversion	Inflammatory Bowel Disease	0.29	0.22	0.18
Extraversion	Insulin	-0.40	0.39	0.31
Extraversion	IQ	-0.30	0.26	0.25
Extraversion	MND	0.74	0.46	0.11
Extraversion	Neuroticism	-1.67	0.79	0.03
Extraversion	Openness	0.40	0.41	0.33
Extraversion	Osteoporosis	-0.14	0.25	0.56
Extraversion	Pubertal Growth	0.01	0.22	0.98
Extraversion	Rheumatoid Arthritis	-0.04	0.23	0.86
Extraversion	Schizophrenia	-0.18	0.15	0.24
Extraversion	Smoking	-0.10	0.22	0.64
Extraversion	Tanner	-0.09	0.42	0.84
Extraversion	Triglycerides	-0.04	0.13	0.75
Extraversion	Ulcerative Colitis	0.24	0.25	0.32
Extraversion	Waist Hip Ratio	0.02	0.16	0.92
Glucose	Head Circumference	-0.01	0.11	0.90
Glucose	Height	-0.04	0.04	0.38
Glucose	Inflammatory Bowel Disease	0.00	0.06	0.94
Glucose	Insulin	0.31	0.10	0.00
Glucose	IQ	0.00	0.10	1.00
Glucose	MND	-0.04	0.13	0.78

Glucose	Neuroticism	0.02	0.10	0.86
Glucose	Openness	-0.10	0.12	0.40
Glucose	Osteoporosis	-0.22	0.07	0.00
Glucose	Pubertal Growth	-0.03	0.07	0.64
Glucose	Rheumatoid Arthritis	0.00	0.07	0.95
Glucose	Schizophrenia	-0.04	0.03	0.25
Glucose	Smoking	0.08	0.07	0.26
Glucose	Tanner	0.02	0.15	0.88
Glucose	Triglycerides	0.09	0.09	0.30
Glucose	Ulcerative Colitis	0.06	0.08	0.44
Glucose	Waist Hip Ratio	0.03	0.05	0.52
Head Circumference	Height	0.26	0.06	0.00
Head Circumference	Inflammatory Bowel Disease	0.04	0.09	0.68
Head Circumference	Insulin	0.01	0.13	0.92
Head Circumference	IQ	0.37	0.16	0.02
Head Circumference	MND	0.02	0.22	0.93
Head Circumference	Neuroticism	-0.05	0.19	0.80
Head Circumference	Openness	0.01	0.21	0.95
Head Circumference	Osteoporosis	-0.15	0.10	0.15
Head Circumference	Pubertal Growth	0.44	0.11	0.00
Head Circumference	Rheumatoid Arthritis	0.02	0.11	0.88
Head Circumference	Schizophrenia	-0.03	0.07	0.62
Head Circumference	Smoking	-0.01	0.11	0.91
Head Circumference	Tanner	0.25	0.23	0.29
Head Circumference	Triglycerides	-0.06	0.06	0.33
Head Circumference	Ulcerative Colitis	0.01	0.12	0.94
Head Circumference	Waist Hip Ratio	-0.17	0.08	0.04
Height	Inflammatory Bowel Disease	0.06	0.03	0.07
Height	Insulin	0.05	0.04	0.24

Height	IQ	0.11	0.05	0.03
Height	MND	-0.07	0.06	0.21
Height	Neuroticism	-0.13	0.05	0.01
Height	Openness	-0.07	0.06	0.24
Height	Osteoporosis	0.00	0.03	0.95
Height	Pubertal Growth	0.75	0.04	0.00
Height	Rheumatoid Arthritis	0.02	0.03	0.46
Height	Schizophrenia	0.01	0.02	0.77
Height	Smoking	-0.06	0.03	0.09
Height	Tanner	-0.03	0.07	0.70
Height	Triglycerides	-0.07	0.02	0.00
Height	Ulcerative Colitis	0.05	0.04	0.22
Height	Waist Hip Ratio	-0.03	0.03	0.26
Inflammatory Bowel Disease	Insulin	0.00	0.06	0.97
Inflammatory Bowel Disease	IQ	-0.11	0.10	0.25
Inflammatory Bowel Disease	MND	0.19	0.11	0.09
Inflammatory Bowel Disease	Neuroticism	-0.02	0.08	0.84
Inflammatory Bowel Disease	Openness	-0.08	0.11	0.44
Inflammatory Bowel Disease	Osteoporosis	0.11	0.05	0.02
Inflammatory Bowel Disease	Pubertal Growth	-0.04	0.06	0.55
Inflammatory Bowel Disease	Rheumatoid Arthritis	0.05	0.06	0.40
Inflammatory Bowel Disease	Schizophrenia	0.14	0.03	0.00

Inflammatory Bowel Disease	Smoking	-0.04	0.06	0.50
Inflammatory Bowel Disease	Tanner	-0.18	0.12	0.14
Inflammatory Bowel Disease	Triglycerides	0.02	0.04	0.57
Inflammatory Bowel Disease	Ulcerative Colitis	0.94	0.02	0.00
Inflammatory Bowel Disease	Waist Hip Ratio	-0.03	0.04	0.48
Insulin	IQ	-0.24	0.11	0.03
Insulin	MND	-0.31	0.19	0.10
Insulin	Neuroticism	-0.02	0.12	0.90
Insulin	Openness	-0.24	0.15	0.10
Insulin	Osteoporosis	-0.09	0.09	0.33
Insulin	Pubertal Growth	0.07	0.09	0.43
Insulin	Rheumatoid Arthritis	0.06	0.08	0.42
Insulin	Schizophrenia	0.02	0.05	0.61
Insulin	Smoking	0.20	0.09	0.03
Insulin	Tanner	0.26	0.20	0.19
Insulin	Triglycerides	0.42	0.09	0.00
Insulin	Ulcerative Colitis	-0.06	0.07	0.46
Insulin	Waist Hip Ratio	0.33	0.07	0.00
IQ	MND	0.00	0.16	1.00
IQ	Neuroticism	0.04	0.17	0.83
IQ	Openness	0.42	0.15	0.00
IQ	Osteoporosis	0.13	0.09	0.14
IQ	Pubertal Growth	0.08	0.09	0.40
IQ	Rheumatoid Arthritis	-0.13	0.09	0.14
IQ	Schizophrenia	-0.05	0.06	0.37
IQ	Smoking	-0.28	0.11	0.01

IQ	Tanner	-0.24	0.24	0.32
IQ	Triglycerides	-0.17	0.06	0.00
IQ	Ulcerative Colitis	-0.01	0.10	0.92
IQ	Waist Hip Ratio	-0.10	0.06	0.07
MND	Neuroticism	0.08	0.20	0.68
MND	Openness	0.14	0.20	0.49
MND	Osteoporosis	-0.04	0.12	0.71
MND	Pubertal Growth	0.04	0.12	0.73
MND	Rheumatoid Arthritis	0.00	0.13	0.99
MND	Schizophrenia	0.25	0.08	0.00
MND	Smoking	0.16	0.14	0.26
MND	Tanner	-0.08	0.23	0.73
MND	Triglycerides	-0.02	0.09	0.84
MND	Ulcerative Colitis	0.11	0.14	0.42
MND	Waist Hip Ratio	-0.02	0.10	0.85
Neuroticism	Openness	0.15	0.20	0.46
Neuroticism	Osteoporosis	0.22	0.10	0.02
Neuroticism	Pubertal Growth	-0.13	0.12	0.29
Neuroticism	Rheumatoid Arthritis	-0.05	0.11	0.65
Neuroticism	Schizophrenia	0.18	0.06	0.00
Neuroticism	Smoking	0.15	0.10	0.15
Neuroticism	Tanner	0.33	0.25	0.19
Neuroticism	Triglycerides	0.21	0.08	0.01
Neuroticism	Ulcerative Colitis	-0.01	0.10	0.88
Neuroticism	Waist Hip Ratio	0.19	0.08	0.02
Openness	Osteoporosis	-0.03	0.10	0.80
Openness	Pubertal Growth	0.08	0.11	0.48
Openness	Rheumatoid Arthritis	0.08	0.11	0.46
Openness	Schizophrenia	0.24	0.08	0.00
Openness	Smoking	-0.13	0.13	0.35
Openness	Tanner	0.57	0.25	0.02

Openness	Triglycerides	-0.05	0.06	0.46
Openness	Ulcerative Colitis	0.02	0.12	0.89
Openness	Waist Hip Ratio	-0.14	0.08	0.07
Osteoporosis	Pubertal Growth	-0.06	0.07	0.39
Osteoporosis	Rheumatoid Arthritis	0.02	0.06	0.71
Osteoporosis	Schizophrenia	0.01	0.04	0.80
Osteoporosis	Smoking	-0.01	0.07	0.87
Osteoporosis	Tanner	0.04	0.13	0.76
Osteoporosis	Triglycerides	0.00	0.04	0.96
Osteoporosis	Ulcerative Colitis	0.05	0.06	0.38
Osteoporosis	Waist Hip Ratio	-0.03	0.05	0.50
Pubertal Growth	Rheumatoid Arthritis	-0.05	0.07	0.52
Pubertal Growth	Schizophrenia	-0.03	0.04	0.42
Pubertal Growth	Smoking	-0.13	0.07	0.08
Pubertal Growth	Tanner	0.42	0.14	0.00
Pubertal Growth	Triglycerides	-0.01	0.04	0.89
Pubertal Growth	Ulcerative Colitis	-0.11	0.07	0.14
Pubertal Growth	Waist Hip Ratio	-0.02	0.05	0.73
Rheumatoid Arthritis	Schizophrenia	0.03	0.04	0.56
Rheumatoid Arthritis	Smoking	0.15	0.06	0.02
Rheumatoid Arthritis	Tanner	-0.05	0.14	0.70
Rheumatoid Arthritis	Triglycerides	-0.02	0.04	0.64
Rheumatoid Arthritis	Ulcerative Colitis	0.10	0.07	0.13
Rheumatoid Arthritis	Waist Hip Ratio	0.02	0.05	0.59
Schizophrenia	Smoking	0.10	0.04	0.01
Schizophrenia	Tanner	0.10	0.08	0.22
Schizophrenia	Triglycerides	-0.02	0.03	0.45
Schizophrenia	Ulcerative Colitis	0.14	0.03	0.00
Schizophrenia	Waist Hip Ratio	0.00	0.03	0.97
Smoking	Tanner	-0.13	0.14	0.33
Smoking	Triglycerides	0.13	0.04	0.00

Smoking	Ulcerative Colitis	-0.05	0.07	0.47
Smoking	Waist Hip Ratio	0.13	0.05	0.00
Tanner	Triglycerides	0.11	0.07	0.13
Tanner	Ulcerative Colitis	-0.21	0.15	0.15
Tanner	Waist Hip Ratio	-0.01	0.10	0.89
Triglycerides	Ulcerative Colitis	-0.01	0.05	0.78
Triglycerides	Waist Hip Ratio	0.32	0.04	0.00
Ulcerative Colitis	Waist Hip Ratio	-0.04	0.04	0.35

Discussion / General conclusion

The aim of this thesis was to elucidate three particular aspects of the genetic architecture of psychiatric disorders: The impact of genetic heterogeneity on heritability estimates (Chapter 2: Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability), the feasibility of detecting genetic heterogeneity through genotype clustering (Chapter 3: Genotype based clustering) and the degree to which the genetic similarity between psychiatric disorders makes it possible to improve genetic risk predictors (Chapter 4: Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder and Chapter 5: Improving genetic prediction by leveraging genetic correlations among human diseases and traits).

Here, I will first review the aims and findings of the chapters centered around genetic heterogeneity and then those of the chapters about multi trait genetic risk prediction.

Genetic heterogeneity

The first heterogeneity related question was to investigate how genetic heterogeneity affects family based heritability estimates compared to how it affects SNP-heritability estimates. If the two are affected differently, heterogeneity can contribute to missing heritability. We assumed an extreme case of zero genetic correlation between two equally sized disease subgroups and found that in this case SNP-heritability estimates will be substantially lower than family based heritability estimates. Even under less extreme assumptions, the presence of genetic heterogeneity can contribute to missing heritability.

The second heterogeneity related question was to investigate the feasibility of genotype based clustering. The relative diagnostic uncertainty in psychiatry makes researchers as well as clinicians look to fields like genetics and brain imaging in the hope for biomarkers that could lead more accurate diagnoses and more individualized interventions. Recently the term “stratified medicine” is often used in this context [111].

Reports of the successful detection of schizophrenia subtypes based on genetic data have

been met with great interest as well as with skepticism [159], but to date no such genotype based sub-classification has withstood the test of time and replication. Motivated by preliminary positive results based on simulated, unstructured genotype data, we set out to explore the conditions under which genotype based clustering can achieve reasonable separation between simulated case subtypes. In our simulation setup we assumed that we know nothing about what separates the subtypes from one another, and that we only have genotype data as well as estimates of the effects that separate the combined case group from the controls. The results were sobering: Only when very few SNPs explained most of the genetic variance, and when we could accurately estimate these effects, the clustering was able to achieve a reasonable separation of cases and controls. Why is that?

In some way, what we attempted to do is similar to polygenic risk prediction: Given information on genotypes and SNP effect estimates, what can we learn about the individual's phenotype? However, there are notable differences: In the clustering case, (i) there is no separate test set, (ii) the SNP effects are only to some degree related to the quantity we want to predict, (iii) the SNP effects are only used to decide which SNP to use for the clustering, (iv) the similarity between individuals determines the predicted phenotype. While (i) should play strongly in our favor, (ii) to (iv) explain why compared to genetic risk predictors, the outcome was disappointing. The curse of dimensionality explains why (iv) is a problem: As the number of markers on which to cluster increases, the variance in the distance among individuals decreases, until it becomes almost meaningless. This is less of a problem in a method like GREML, which is based on the pairwise similarity among individuals, because there, a large number of pairwise similarity measures are used to estimate a small number of parameters, and not to accurately group a large number of individuals.

The small overlap between clusters and simulated sub-types didn't leave much hope for this approach to be successful in practice, especially because a number of optimistic assumptions were made in the simulations that are not likely to hold in practice: (i) whereas the simulations assume distinct and well defined groups, a continuum of different genetic risk profiles seems more probable. Shades of grey are more common than black and white. (ii) even if distinct and discrete genetic clusters exist, they are unlikely to be equal in size and there are probably more than just two of them.

Given that even under these simplifying assumptions the clustering could only recapture sub-types when the number of markers explaining most of the variance was very small, it appears more promising to pursue other paths. For example, the simpler problem of detecting the presence of genetic heterogeneity seems to be more tractable [112], as is the task of detecting sub-groups by combining both genetic and non-genetic data. These non-genetic data could include traditional phenotypes, for example environmental covariates, or molecular phenotypes such as gene expression, DNA methylation, or other epigenetic modifications.

Multi trait risk prediction

Due to the similarity between the two multi trait risk prediction projects, they will here be discussed together.

Aims and findings

The first goal of the multi trait risk prediction projects was to develop methodology that allows to combine data on multiple traits such that genetic risk predictors for each of the traits will be more accurate than corresponding single trait predictors. The second goal was to test empirically how much such a multi trait risk predictor increases accuracy.

In the first study, based on individual level genotype data, the method was tested on five psychiatric disorders using PGC wave 1 data and found consistent improvements, both in a cross validation design, as well as in independent samples for three of the disorders. The gain was benchmarked as equivalence in sample size increases between 34% and 76%.
In the second, summary statistics based projects, the method was first tested on two out of the five traits from the first project, and achieved very similar prediction accuracy. Further tests on different validation cohorts complicated the picture. Application to a wide range of other traits resulted in consistent improvements over single trait predictors.

Differences between the two multi trait projects

Apart from the main distinction of individual level data versus summary statistics data, there are several other differences between the two projects that are worth discussing. Some of these differences were imposed by the data analyzed, others just reflect a focus on different aspects of the analysis.

Summary data versus individual level data. The motivation behind the second project was to replicate the results of the first projects using summary statistics data, with the previously discussed advantages related to computation and ease of access. One implication of using summary data was that we could show that even for non-psychiatric traits, where genetic correlations are generally lower, this method can provide a benefit. Another implication of using summary data is that what was one step in the first project, became two separate steps in the second project (SBLUP transformation and weighting of multiple traits). There is some loss of accuracy associated with both of these steps, arising from the use of summary statistics. These are discussed separately further below.

Prediction accuracy differences. A central question in both projects was to find out if the prediction from one method is more accurate than the prediction from another method. In the first project, the difference in accuracy between two predictors was evaluated by a likelihood ratio test of two nested models: The first model contained only one predictor (the single trait predictor), and the second model contained both predictors (single trait and multi trait). If the difference between the two models was significant, we deemed the multi trait predictor to be significantly better. This procedure is valid, but it treats the validation set as fixed.

If, on the other hand, the validation set is seen as a random sample from a larger population, it is desirable to account for this source of variation. This can be done by calculating the standard error of the correlation coefficient, or similarly by splitting the validation data into random subsets. Establishing a significant difference between two prediction methods in this case requires a larger sample size in the validation set. The UK Biobank provided us with a large enough data set to test significance in this more stringent way and still find significant differences.

The third way to test the significance of the difference between two predictors is by treating the validation set as a *non*-random sample of a larger population. In this case there is even more variation among the validation data, which means that large validation set sample sizes and large prediction accuracy differences would be needed to establish significance. On the other hand, this method provides the highest confidence in the generalizability of the result and the best protection against labelling a difference as significant that is in fact just a consequence of overfitting.

The cohort-wise summary statistics in the PGC2 data provided us with the opportunity to compare predictors in this most stringent way. However, due to the large variation between cohorts and the limited number of cohorts, this leave-one-cohort out validation procedure did not allow us to find significant differences between the predictors. It did, however, open the door for further analyses which allowed to explore the impact of heterogeneity on the prediction results.

Cohort wise analysis. In the PGC wave 2 schizophrenia and bipolar disorder analysis we could use each cohort as a validation set, while using the other cohorts as training sets. This highlighted the substantial amount of heterogeneity among cohorts, as the prediction accuracies varied substantially between the different validation cohorts. In some cases, the relative accuracy of single trait and multi trait predictors could be explained by different ascertainment of cases in the validation cohorts.

Exploring the parameter space. In the first project, we tested the impact of inclusion or exclusion of specific traits on the prediction accuracy. The low computational runtime of the method developed in the second project allowed us to go further than that and to test the impact of gradually increasing or decreasing the contribution of specific traits on the prediction accuracy. In the comparison of schizophrenia and bipolar disorder this revealed that when adding schizophrenia data to bipolar data, the weights were on average close to optimal, but in the other direction the weights we used were too much skewed towards bipolar disorder. This could be a consequence of an overestimation of SNP-heritability of bipolar disorder and of the genetic correlation in this data set.

Impact of population outliers on prediction. In the first project we went to great lengths to show that the prediction results are not driven by population outliers. Having found that population outliers don't have a large effect, in the second project we settled for a standard principal component correction.

Notation. Finally, the equations in both projects were parameterized differently. For example, Equation (3) in Chapter 4 is equivalent to Equation (9) in Chapter 5, even though this may not be immediately obvious. See Appendix section "Comparison of different BLUP formulations".

Limitations

Sample overlap. In contrast to the first project, the method developed in the second project assumes that the data that are to be combined into a multi trait predictor were collected on non-overlapping individuals. The extent of sample overlap can be estimated from summary statistics [9], but the model developed here does not account for overlap. Consequently, the weights may not be optimal when traits are combined that are based on overlapping samples.

Differences in the number of markers. For many GWAS summary statistics data sets, the number of individuals is more or less constant across SNPs, apart from small fluctuations caused by different rates of missingness in the QC process. In some meta analyzed sets however, there can be larger differences between SNPs. As our method uses the median sample size across all markers, and assigns one constant weight to each trait, large differences in the number of individuals within a data set can also lead to weights that don't maximize prediction accuracy.

Treating heritability and genetic correlation estimates as fixed. Both methods of multi trait prediction rely on accurate estimates of SNP heritability and genetic correlation. In the method using individual level genotype data GREML is used to estimate these parameters, whereas the summary statistics based method uses LD score regression, which has larger standard errors. However, our expectations for the multi trait prediction accuracy assume that these parameters have been estimated without error. This is especially problematic when using data with low sample size as this increases the standard error, and when combining many traits. To protect against that, we applied a threshold on the genetic

correlation p-value estimates when choosing traits to combine in the application of the summary statistics based method.

Limitations of SBLUP. BLUP is a well-established prediction method, but the concept of approximating BLUP from summary statistics is relatively new. While under ideal conditions (perfect match of the reference set and large window size) SBLUP is identical to BLUP, under more realistic conditions SBLUP will not be as accurate as BLUP. Furthermore, our simulations have revealed that when an external reference data set is used to approximate the LD structure, the optimal shrinkage factor increases, but there is currently no theory that describes the magnitude of this effect.

Biased predictors. A multi trait predictor for bipolar disorder that incorporates data from schizophrenia will perform better in individuals who have both a high liability for bipolar disorder and schizophrenia, than in individuals who have a high liability for bipolar disorder, but a low liability for schizophrenia. As such, it is not a “pure” bipolar disorder predictor anymore. We don’t view this as a major limitation, since the key metric of a predictor is its overall accuracy. It is for this reasons, however, that we decided to apply the methodology presented here only to prediction, and not to the discovery of associated variants: A SNP that is significantly associated with bipolar disorder, because it has the same direction of effect in schizophrenia and bipolar disorder, has to be interpreted differently than a SNP which is significant because of its direction of effect in bipolar disorder, independent of its effect on schizophrenia.

Overall conclusions

The main contributions of this thesis are to highlight the impact of genetic heterogeneity on estimates of heritability, and the development and application of methods for multi trait genetic risk prediction. Even after harnessing the combined effect of multiple traits, the prediction accuracy for many traits is still too low for many potential applications, but in combination with other improvements to risk prediction as well as with growing sample sizes, accuracy will eventually be high enough to allow accurate genetic predictions of many traits and diseases. When that will be the case, polygenic risk prediction will not only be a widely used research method, but a powerful tool for preventing disease.

Bibliography

[1] K.S. Kendler, Psychiatric genetics: A methodologic critique, Am. J. Psychiatry. 162 (2005) 3–11. doi:10.1176/appi.ajp.162.1.3.

[2] K.S. Kendler, What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn., Mol. Psychiatry. 18 (2013) 1058–66. doi:10.1038/mp.2013.50.

[3] P.F. Sullivan, M.J. Daly, M. O'Donovan, Genetic architectures of psychiatric disorders: the emerging picture and its implications., Nat. Rev. Genet. 13 (2012) 537–51. doi:10.1038/nrg3240.

[4] B. Pasaniuc, A.L. Price, Dissecting the genetics of complex traits using summary association statistics, Nat. Rev. Genet. 18 (2016) 117–127. doi:10.1038/nrg.2016.142.

[5] K.S. Kendler, Toward a Scientific Psychiatric Nosology, Arch. Gen. Psychiatry. 47 (1990) 969. doi:10.1001/archpsyc.1990.01810220085011.

[6] E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, J.H. Nadeau, Missing heritability and strategies for finding the underlying causes of complex disease., Nat. Rev. Genet. 11 (2010) 446–50. doi:10.1038/nrg2809.

[7] T. a Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L. a Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F.C. Mackay, S. a McCarroll, P.M. Visscher, Finding the missing heritability of complex diseases., Nature. 461 (2009) 747–53. doi:10.1038/nature08494.

[8] J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. a E. Vinkhuyzen, S.H. Lee, M.R. Robinson, J.R.B. Perry, I.M. Nolte, J. V van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P.K.E. Magnusson, N.L. Pedersen, E. Ingelsson, N. Soranzo, M.C. Keller, N.R. Wray, M.E. Goddard, P.M. Visscher, Genetic variance estimation with imputed variants finds negligible missing heritability for human height

and body mass index, Nat. Genet. 47 (2015) 1114–1120. doi:10.1038/ng.3390.

[9] B. Bulik-Sullivan, H.K. Finucane, V. Anttila, A. Gusev, F.R. Day, P.-R. Loh, L. Duncan, J.R.B. Perry, N. Patterson, E.B. Robinson, M.J. Daly, A.L. Price, B.M. Neale, An atlas of genetic correlations across human diseases and traits., Nat. Genet. 47 (2015) 1236–1241. doi:10.1038/ng.3406.

[10] S.H. Lee, S. Ripke, B.M. Neale, S. V Faraone, S.M. Purcell, R.H. Perlis, B.J. Mowry, A. Thapar, M.E. Goddard, J.S. Witte, D. Absher, I. Agartz, H. Akil, F. Amin, O. a Andreassen, A. Anjorin, R. Anney, V. Anttila, D.E. Arking, P. Asherson, M.H. Azevedo, L. Backlund, J. a Badner, A.J. Bailey, T. Banaschewski, J.D. Barchas, M.R. Barnes, T.B. Barrett, N. Bass, A. Battaglia, M. Bauer, M. Bayés, F. Bellivier, S.E. Bergen, W. Berrettini, C. Betancur, T. Bettecken, J. Biederman, E.B. Binder, D.W. Black, D.H.R. Blackwood, C.S. Bloss, M. Boehnke, D.I. Boomsma, G. Breen, R. Breuer, R. Bruggeman, P. Cormican, N.G. Buccola, J.K. Buitelaar, W.E. Bunney, J.D. Buxbaum, W.F. Byerley, E.M. Byrne, S. Caesar, W. Cahn, R.M. Cantor, M. Casas, A. Chakravarti, K. Chambert, K. Choudhury, S. Cichon, C.R. Cloninger, D. a Collier, E.H. Cook, H. Coon, B. Cormand, A. Corvin, W.H. Coryell, D.W. Craig, I.W. Craig, J. Crosbie, M.L. Cuccaro, D. Curtis, D. Czamara, S. Datta, G. Dawson, R. Day, E.J. De Geus, F. Degenhardt, S. Djurovic, G.J. Donohoe, A.E. Doyle, J. Duan, F. Dudbridge, E. Duketis, R.P. Ebstein, H.J. Edenberg, J. Elia, S. Ennis, B. Etain, A. Fanous, A.E. Farmer, I.N. Ferrier, M. Flickinger, E. Fombonne, T. Foroud, J. Frank, B. Franke, C. Fraser, R. Freedman, N.B. Freimer, C.M. Freitag, M. Friedl, L. Frisén, L. Gallagher, P. V Gejman, L. Georgieva, E.S. Gershon, D.H. Geschwind, I. Giegling, M. Gill, S.D. Gordon, K. Gordon-Smith, E.K. Green, T. a Greenwood, D.E. Grice, M. Gross, D. Grozeva, W. Guan, H. Gurling, L. De Haan, J.L. Haines, H. Hakonarson, J. Hallmayer, S.P. Hamilton, M.L. Hamshere, T.F. Hansen, A.M. Hartmann, M. Hautzinger, A.C. Heath, A.K. Henders, S. Herms, I.B. Hickie, M. Hipolito, S. Hoefels, P. a Holmans, F. Holsboer, W.J. Hoogendijk, J.-J. Hottenga, C.M. Hultman, V. Hus, A. Ingason, M. Ising, S. Jamain, E.G. Jones, I. Jones, L. Jones, J.-Y. Tzeng, A.K. Kähler, R.S. Kahn, R. Kandaswamy, M.C. Keller, J.L. Kennedy, E. Kenny, L. Kent, Y. Kim, G.K. Kirov, S.M. Klauck, L. Klei, J. a Knowles, M. a Kohli, D.L. Koller, B. Konte, A. Korszun, L. Krabbendam, R. Krasucki, J. Kuntsi, P. Kwan, M. Landén, N. Långström, M. Lathrop, J. Lawrence, W.B. Lawson, M. Leboyer, D.H. Ledbetter, P.H. Lee, T. Lencz, K.-P. Lesch, D.F. Levinson, C.M. Lewis, J. Li, P. Lichtenstein, J. a Lieberman, D.-Y. Lin,

D.H. Linszen, C. Liu, F.W. Lohoff, S.K. Loo, C. Lord, J.K. Lowe, S. Lucae, D.J. MacIntyre, P. a F. Madden, E. Maestrini, P.K.E. Magnusson, P.B. Mahon, W. Maier, A.K. Malhotra, S.M. Mane, C.L. Martin, N.G. Martin, M. Mattheisen, K. Matthews, M. Mattingsdal, S. a McCarroll, K. a McGhee, J.J. McGough, P.J. McGrath, P. McGuffin, M.G. McInnis, A. McIntosh, R. McKinney, A.W. McLean, F.J. McMahon, W.M. McMahon, A. McQuillin, H. Medeiros, S.E. Medland, S. Meier, I. Melle, F. Meng, J. Meyer, C.M. Middeldorp, L. Middleton, V. Milanova, A. Miranda, A.P. Monaco, G.W. Montgomery, J.L. Moran, D. Moreno-De-Luca, G. Morken, D.W. Morris, E.M. Morrow, V. Moskvina, P. Muglia, T.W. Mühleisen, W.J. Muir, B. Müller-Myhsok, M. Murtha, R.M. Myers, I. Myin-Germeys, M.C. Neale, S.F. Nelson, C.M. Nievergelt, I. Nikolov, V. Nimgaonkar, W. a Nolen, M.M. Nöthen, J.I. Nurnberger, E. a Nwulia, D.R. Nyholt, C. O'Dushlaine, R.D. Oades, A. Olincy, G. Oliveira, L. Olsen, R. a Ophoff, U. Osby, M.J. Owen, A. Palotie, J.R. Parr, A.D. Paterson, C.N. Pato, M.T. Pato, B.W. Penninx, M.L. Pergadia, M. a Pericak-Vance, B.S. Pickard, J. Pimm, J. Piven, D. Posthuma, J.B. Potash, F. Poustka, P. Propping, V. Puri, D.J. Quested, E.M. Quinn, J.A. Ramos-Quiroga, H.B. Rasmussen, S. Raychaudhuri, K. Rehnström, A. Reif, M. Ribasés, J.P. Rice, M. Rietschel, K. Roeder, H. Roeyers, L. Rossin, A. Rothenberger, G. Rouleau, D. Ruderfer, D. Rujescu, A.R. Sanders, S.J. Sanders, S.L. Santangelo, J. a Sergeant, R. Schachar, M. Schalling, A.F. Schatzberg, W. a Scheftner, G.D. Schellenberg, S.W. Scherer, N.J. Schork, T.G. Schulze, J. Schumacher, M. Schwarz, E. Scolnick, L.J. Scott, J. Shi, P.D. Shilling, S.I. Shyn, J.M. Silverman, S.L. Slager, S.L. Smalley, J.H. Smit, E.N. Smith, E.J.S. Sonuga-Barke, D. St. Clair, M. State, M. Steffens, H.-C. Steinhausen, J.S. Strauss, J. Strohmaier, T.S. Stroup, J.S. Sutcliffe, P. Szatmari, S. Szelinger, S. Thirumalai, R.C. Thompson, A. a Todorov, F. Tozzi, J. Treutlein, M. Uhr, E.J.C.G. van den Oord, G. Van Grootheest, J. Van Os, A.M. Vicente, V.J. Vieland, J.B. Vincent, P.M. Visscher, C. a Walsh, T.H. Wassink, S.J. Watson, M.M. Weissman, T. Werge, T.F. Wienker, E.M. Wijsman, G. Willemsen, N. Williams, a J. Willsey, S.H. Witt, W. Xu, A.H. Young, T.W. Yu, S. Zammit, P.P. Zandi, P. Zhang, F.G. Zitman, S. Zöllner, B. Devlin, J.R. Kelsoe, P. Sklar, M.J. Daly, M.C. O'Donovan, N. Craddock, P.F. Sullivan, J.W. Smoller, K.S. Kendler, N.R. Wray, Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs, Nat. Genet. 45 (2013) 984–994. doi:10.1038/ng.2711.

[11] P.M. Visscher, M.A. Brown, M.I. McCarthy, J. Yang, Five years of GWAS discovery., Am. J. Hum. Genet. 90 (2012) 7–24. doi:10.1016/j.ajhg.2011.11.029.

[12] J. Gratten, Rare variants are common in schizophrenia, Nat. Neurosci. 19 (2016) 1426–1428. doi:10.1038/nn.4422.

[13] K. Wang, M. Li, H. Hakonarson, Analysing biological pathways in genome-wide association studies, Nat Rev Genet. 11 (2010) 843–854. doi:10.1038/nrg2884.

[14] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D.A. Stephan, S.F. Nelson, D.W. Craig, Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays, PLoS Genet. 4 (2008) e1000167. doi:10.1371/journal.pgen.1000167.

[15] S. Sankararaman, G. Obozinski, M.I. Jordan, E. Halperin, Genomic privacy and limits of individual detection in a pool, Nat. Genet. 41 (2009) 965–967. doi:10.1038/ng.436.

[16] P.M. Visscher, W.G. Hill, The limits of individual identification from sample allele frequencies: Theory and statistical analysis, PLoS Genet. 5 (2009) 1–6. doi:10.1371/journal.pgen.1000628.

[17] A. a E. Vinkhuyzen, N.R. Wray, J. Yang, M.E. Goddard, P.M. Visscher, Estimation and partition of heritability in human populations using whole-genome analysis methods., Annu. Rev. Genet. 47 (2013) 75–95. doi:10.1146/annurev-genet-111212-133258.

[18] J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P. a Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, P.M. Visscher, Common SNPs explain a large proportion of the heritability for human height., Nat. Genet. 42 (2010) 565–9. doi:10.1038/ng.608.

[19] S.H. Lee, T.R. DeCandia, S. Ripke, J. Yang, P.F. Sullivan, M.E. Goddard, M.C. Keller, P.M. Visscher, N.R. Wray, Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs., Nat. Genet. 44 (2012) 247–50. doi:10.1038/ng.1108.

[20] E.A. Stahl, D. Wegmann, G. Trynka, J. Gutierrez-Achury, R. Do, B.F. Voight, P. Kraft, R. Chen, H.J. Kallberg, F.A.S. Kurreeman, S. Kathiresan, C. Wijmenga, P.K. Gregersen, L. Alfredsson, K.A. Siminovitch, J. Worthington, P.I.W. de Bakker, S. Raychaudhuri, R.M. Plenge, Bayesian inference analyses of the polygenic

architecture of rheumatoid arthritis., Nat. Genet. 44 (2012) 483–9. doi:10.1038/ng.2232.

[21] S.H. Lee, N.R. Wray, M.E. Goddard, P.M. Visscher, Estimating missing heritability for disease from genome-wide association studies., Am. J. Hum. Genet. 88 (2011) 294–305. doi:10.1016/j.ajhg.2011.02.002.

[22] W.J. Peyrot, D.I. Boomsma, B.W.J.H. Penninx, N.R. Wray, Disease and Polygenic Architecture: Avoid Trio Design and Appropriately Account for Unscreened Control Subjects for Common Disease, Am. J. Hum. Genet. 98 (2016) 382–391. doi:10.1016/j.ajhg.2015.12.017.

[23] D. Golan, E.S. Lander, S. Rosset, Measuring missing heritability: Inferring the contribution of common variants., Proc. Natl. Acad. Sci. U. S. A. (2014). doi:10.1073/pnas.1419064111.

[24] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829. doi:11290733.

[25] B.J. Hayes, P.M. Visscher, M.E. Goddard, Increased accuracy of artificial selection by using the realized relationship matrix., Genet. Res. (Camb). 91 (2009) 47–60. doi:10.1017/S0016672308009981.

[26] J. Yang, S.H. Lee, M.E. Goddard, P.M. Visscher, GCTA: a tool for genome-wide complex trait analysis., Am. J. Hum. Genet. 88 (2011) 76–82. doi:10.1016/j.ajhg.2010.11.011.

[27] D. Speed, G. Hemani, M.R. Johnson, D.J. Balding, Improved Heritability Estimation from Genome-wide SNPs, Am. J. Hum. Genet. 91 (2012) 1011–1021. doi:10.1016/j.ajhg.2012.10.010.

[28] X. Zhou, M. Stephens, Efficient multivariate linear mixed model algorithms for genome-wide association studies., Nat. Methods. 11 (2014) 407–9. doi:10.1038/nmeth.2848.

[29] L. Evans, R. Tahmasbi, S. Vrieze, G. Abecasis, S. Das, D. Bjelland, T. DeCandia, H.R. Consortium, M. Goddard, B. Neale, J. Yang, P. Visscher, M. Keller, Comparison of methods that use whole genome data to estimate the heritability and genetic

architecture of complex traits., bioRxiv. (2017) 115527. doi:10.1101/115527.

[30] P.-R. Loh, G. Tucker, B.K. Bulik-Sullivan, B.J. Vilhjálmsson, H.K. Finucane, R.M. Salem, D.I. Chasman, P.M. Ridker, B.M. Neale, B. Berger, N. Patterson, A.L. Price, Efficient Bayesian mixed-model analysis increases association power in large cohorts., Nat. Genet. (2015). doi:10.1038/ng.3190.

[31] J. Yang, N. a Zaitlen, M.E. Goddard, P.M. Visscher, A.L. Price, Advantages and pitfalls in the application of mixed-model association methods., Nat. Genet. 46 (2014) 100–6. doi:10.1038/ng.2876.

[32] B.K. Bulik-Sullivan, P.-R. Loh, H.K. Finucane, S. Ripke, J. Yang, N. Patterson, M.J. Daly, A.L. Price, B.M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, Nat. Genet. 47 (2015) 291–295. doi:10.1038/ng.3211.

[33] J. Yang, M.N. Weedon, S. Purcell, G. Lettre, K. Estrada, C.J. Willer, A. V Smith, E. Ingelsson, J.R. O'Connell, M. Mangino, R. Mägi, P.A. Madden, A.C. Heath, D.R. Nyholt, N.G. Martin, G.W. Montgomery, T.M. Frayling, J.N. Hirschhorn, M.I. McCarthy, M.E. Goddard, P.M. Visscher, Genomic inflation factors under polygenic inheritance., Eur. J. Hum. Genet. 19 (2011) 807–812. doi:10.1038/ejhg.2011.39.

[34] B.K. Bulik-Sullivan, P.-R. Loh, H.K. Finucane, S. Ripke, J. Yang, N. Patterson, M.J. Daly, A.L. Price, B.M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, Nat. Genet. 47 (2015) 291–295. doi:10.1038/ng.3211.

[35] P.M. Visscher, G. Hemani, A. a E. Vinkhuyzen, G.-B. Chen, S.H. Lee, N.R. Wray, M.E. Goddard, J. Yang, Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples., PLoS Genet. 10 (2014) e1004269. doi:10.1371/journal.pgen.1004269.

[36] R. Schweiger, S. Kaufman, R. Laaksonen, M.E. Kleber, W. März, E. Eskin, S. Rosset, E. Halperin, Fast and Accurate Construction of Confidence Intervals for Heritability, Am. J. Hum. Genet. 98 (2016) 1181–1192. doi:10.1016/j.ajhg.2016.04.016.

[37] J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A.A.E. Vinkhuyzen, I.M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P.K.E.

Magnusson, N.L. Pedersen, E. Ingelsson, P.M. Visscher, Genome-wide genetic homogeneity between sexes and populations for human height and body mass index, Hum. Mol. Genet. 24 (2015) 7445–7449. doi:10.1093/hmg/ddv443.

[38] B.C. Brown, C.J. Ye, A.L. Price, N. Zaitlen, Transethnic Genetic-Correlation Estimates from Summary Statistics, Am. J. Hum. Genet. 99 (2016) 76–88. doi:10.1016/j.ajhg.2016.05.001.

[39] A. V Anttila, H. Finucane, J. Bras, L. Duncan, G. Falcone, P. Gormley, R. Malik, N. Patsopoulos, S. Ripke, R. Walters, D. Yu, P.H. Lee, I. Consortium, Analysis of shared heritability in common disorders of the brain Brainstorm consortium, bioRxiv. (2016) 48991. doi:10.1101/048991.

[40] L. Klei, S.J. Sanders, M.T. Murtha, V. Hus, J.K. Lowe, A.J. Willsey, D. Moreno-De-Luca, T.W. Yu, E. Fombonne, D. Geschwind, D.E. Grice, D.H. Ledbetter, C. Lord, S.M. Mane, C.L. Martin, D.M. Martin, E.M. Morrow, C. a Walsh, N.M. Melhem, P. Chaste, J.S. Sutcliffe, M.W. State, E.H.J. Cook, K. Roeder, B. Devlin, Common genetic variants, acting additively, are a major source of risk for autism, Mol. Autism. 3 (2012) 9. doi:10.1186/2040-2392-3-9.

[41] T. Gaugler, L. Klei, S.J. Sanders, C.A. Bodea, A.P. Goldberg, A.B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, S. Ripke, S. Sandin, P. Sklar, O. Svantesson, A. Reichenberg, C.M. Hultman, B. Devlin, K. Roeder, J.D. Buxbaum, Most genetic risk for autism resides with common variation., Nat. Genet. 46 (2014) 881–5. doi:10.1038/ng.3039.

[42] E.B. Robinson, B. St Pourcain, V. Anttila, J.A. Kosmicki, B. Bulik-Sullivan, J. Grove, J. Maller, K.E. Samocha, S.J. Sanders, S. Ripke, J. Martin, M. V Hollegaard, T. Werge, D.M. Hougaard, iPSYCH-SSI-Broad Autism Group, B.M. Neale, D.M. Evans, D. Skuse, P.B. Mortensen, A.D. B\orglum, A. Ronald, G.D. Smith, M.J. Daly, Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population, Nat. Genet. 48 (2015) 552–555. doi:10.1101/027771.

[43] M.-T. Lo, D.A. Hinds, J.Y. Tung, C. Franz, C.-C. Fan, Y. Wang, O.B. Smeland, A. Schork, D. Holland, K. Kauppi, N. Sanyal, V. Escott-Price, D.J. Smith, M. O'Donovan, H. Stefansson, G. Bjornsdottir, T.E. Thorgeirsson, K. Stefansson, L.K. McEvoy, A.M. Dale, O.A. Andreassen, C.-H. Chen, Genome-wide analyses for personality traits

identify six genomic loci and show correlations with psychiatric disorders, Nat. Genet. 49 (2016) 152–156. doi:10.1038/ng.3736.

[44] A. Okbay, B.M.L. Baselmans, J.-E. De Neve, P. Turley, M.G. Nivard, M.A. Fontana, S.F.W. Meddens, R.K. Linnér, C.A. Rietveld, J. Derringer, J. Gratten, J.J. Lee, J.Z. Liu, R. de Vlaming, T.S. Ahluwalia, J. Buchwald, A. Cavadino, A.C. Frazier-Wood, N.A. Furlotte, V. Garfield, M.H. Geisel, J.R. Gonzalez, S. Haitjema, R. Karlsson, S.W. van der Laan, K.-H. Ladwig, J. Lahti, S.J. van der Lee, P.A. Lind, T. Liu, L. Matteson, E. Mihailov, M.B. Miller, C.C. Minica, I.M. Nolte, D. Mook-Kanamori, P.J. van der Most, C. Oldmeadow, Y. Qian, O. Raitakari, R. Rawal, A. Realo, R. Rueedi, B. Schmidt, A. V Smith, E. Stergiakouli, T. Tanaka, K. Taylor, J. Wedenoja, J. Wellmann, H.-J. Westra, S.M. Willems, W. Zhao, N. Amin, A. Bakshi, P.A. Boyle, S. Cherney, S.R. Cox, G. Davies, O.S.P. Davis, J. Ding, N. Direk, P. Eibich, R.T. Emeny, G. Fatemifar, J.D. Faul, L. Ferrucci, A. Forstner, C. Gieger, R. Gupta, T.B. Harris, J.M. Harris, E.G. Holliday, J.-J. Hottenga, P.L. De Jager, M.A. Kaakinen, E. Kajantie, V. Karhunen, I. Kolcic, M. Kumari, L.J. Launer, L. Franke, R. Li-Gao, M. Koini, A. Loukola, P. Marques-Vidal, G.W. Montgomery, M.A. Mosing, L. Paternoster, A. Pattie, K.E. Petrovic, L. Pulkki-Råback, L. Quaye, K. Räikkönen, I. Rudan, R.J. Scott, J.A. Smith, A.R. Sutin, M. Trzaskowski, A.E. Vinkhuyzen, L. Yu, D. Zabaneh, J.R. Attia, D.A. Bennett, K. Berger, L. Bertram, D.I. Boomsma, H. Snieder, S.-C. Chang, F. Cucca, I.J. Deary, C.M. van Duijn, J.G. Eriksson, U. Bültmann, E.J.C. de Geus, P.J.F. Groenen, V. Gudnason, T. Hansen, C.A. Hartman, C.M.A. Haworth, C. Hayward, A.C. Heath, D.A. Hinds, E. Hyppönen, W.G. Iacono, M.-R. Järvelin, K.-H. Jöckel, J. Kaprio, S.L.R. Kardia, L. Keltikangas-Järvinen, P. Kraft, L.D. Kubzansky, T. Lehtimäki, P.K.E. Magnusson, N.G. Martin, M. McGue, A. Metspalu, M. Mills, R. de Mutsert, A.J. Oldehinkel, G. Pasterkamp, N.L. Pedersen, R. Plomin, O. Polasek, C. Power, S.S. Rich, F.R. Rosendaal, H.M. den Ruijter, D. Schlessinger, H. Schmidt, R. Svento, R. Schmidt, B.Z. Alizadeh, T.I.A. Sørensen, T.D. Spector, A. Steptoe, A. Terracciano, A.R. Thurik, N.J. Timpson, H. Tiemeier, A.G. Uitterlinden, P. Vollenweider, G.G. Wagner, D.R. Weir, J. Yang, D.C. Conley, G.D. Smith, A. Hofman, M. Johannesson, D.I. Laibson, S.E. Medland, M.N. Meyer, J.K. Pickrell, T. Esko, R.F. Krueger, J.P. Beauchamp, P.D. Koellinger, D.J. Benjamin, M. Bartels, D. Cesarini, Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses, Nat. Genet. 48 (2016) 624–633. doi:10.1038/ng.3552.

[45] H.K. Finucane, B.K. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F.R. Day, S. Purcell, E. Stahl, S. Lindstrom, J.R.B. Perry, Y. Okada, S. Raychaudhuri, M.J. Daly, N. Patterson, B.M. Neale, A.L. Price, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Partitioning heritability by functional annotation using genome-wide association summary statistics, Nat. Genet. 47 (2015) 1228–1235. doi:10.1038/ng.3404.

[46] K.S. Kendler, The impact of diagnostic misclassification on the pattern of familial aggregation and coaggregation of psychiatric illness., J. Psychiatr. Res. 21 (1987) 55–91. doi:http://dx.doi.org/10.1016/0022-3956(87)90008-2.

[47] N.R. Wray, S.H. Lee, K.S. Kendler, Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes., Eur. J. Hum. Genet. 20 (2012) 668–74. doi:10.1038/ejhg.2011.257.

[48] S.H. Lee, J. Yang, M.E. Goddard, P.M. Visscher, N.R. Wray, Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood., Bioinformatics. 28 (2012) 2540–2. doi:10.1093/bioinformatics/bts474.

[49] J. Zheng, A.M. Erzurumluoglu, B.L. Elsworth, J.P. Kemp, L. Howe, P.C. Haycock, G. Hemani, K. Tansey, C. Laurin, Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, B.S. Pourcain, N.M. Warrington, H.K. Finucane, A.L. Price, B.K. Bulik-Sullivan, V. Anttila, L. Paternoster, T.R. Gaunt, D.M. Evans, B.M. Neale, 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 (2017) 272–279. doi:10.1093/bioinformatics/btw613.

[50] T.R. De Candia, S.H. Lee, J. Yang, B.L. Browning, P. V. Gejman, D.F. Levinson, B.J. Mowry, J.K. Hewitt, M.E. Goddard, M.C. O'Donovan, S.M. Purcell, D. Posthuma, P.M. Visscher, N.R. Wray, M.C. Keller, Additive genetic variation in schizophrenia risk is shared by populations of African and European descent, Am. J. Hum. Genet. 93 (2013) 463–470. doi:10.1016/j.ajhg.2013.07.007.

[51] N.R. Wray, J. Yang, M.E. Goddard, P.M. Visscher, The genetic interpretation of area

under the ROC curve in genomic profiling., PLoS Genet. 6 (2010) e1000864. doi:10.1371/journal.pgen.1000864.

[52] P.R. Joyce, Age of onset in bipolar affective disorder and misdiagnosis as schizophrenia., Psychol. Med. 14 (1984) 145–149. doi:10.1017/S0033291700003147.

[53] F. Meyer, T.D. Meyer, The misdiagnosis of bipolar disorder as a psychotic disorder: Some of its causes and their influence on therapy, J. Affect. Disord. 112 (2009) 174–183. doi:10.1016/j.jad.2008.04.022.

[54] N. Wray, M. Goddard, P. Visscher, Prediction of individual genetic risk to disease from genome-wide association studies, Genome Res. 17 (2007) 1520–1528. doi:10.1101/gr.6665407.1520.

[55] S.M. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, P.F. Sullivan, P. Sklar, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder., Nature. 460 (2009) 748–52. doi:10.1038/nature08185.

[56] N. Chatterjee, J. Shi, M. García-Closas, Developing and evaluating polygenic risk prediction models for stratified disease prevention., Nat. Rev. Genet. 14210 (2016) 14205–14210. doi:10.1038/nrg.2016.27.

[57] F. Dudbridge, Power and predictive accuracy of polygenic risk scores., PLoS Genet. 9 (2013) e1003348. doi:10.1371/journal.pgen.1003348.

[58] S. Ripke, B.M. Neale, A. Corvin, J.T.R. Walters, K.-H. Farh, P. a. Holmans, P. Lee, B. Bulik-Sullivan, D. a. Collier, H. Huang, T.H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. a. Bacanu, M. Begemann, R. a. Belliveau Jr, J. Bene, S.E. Bergen, E. Bevilacqua, T.B. Bigdeli, D.W. Black, R. Bruggeman, N.G. Buccola, R.L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R.M. Cantor, V.J. Carr, N. Carrera, S. V. Catts, K.D. Chambert, R.C.K. Chan, R.Y.L. Chen, E.Y.H. Chen, W. Cheng, E.F.C. Cheung, S. Ann Chong, C. Robert Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J.J. Crowley, D. Curtis, M. Davidson, K.L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A.H. Fanous, M.S. Farrell, J. Frank, L. Franke,

R. Freedman, N.B. Freimer, M. Friedl, J.I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J.I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M.L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A.M. Hartmann, F. a. Henskens, S. Herms, J.N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D.M. Hougaard, M. Ikeda, I. Joa, A. Julià, R.S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M.C. Keller, J.L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. a. Knowles, B. Konte, V. Kucinskas, Z. Ausrele Kucinskiene, H. Kuzelova-Ptackova, A.K. Kähler, C. Laurent, J. Lee Chee Keong, S. Hong Lee, S.E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C.M. Loughland, J. Lubinski, J. Lönnqvist, M. Macek Jr, P.K.E. Magnusson, B.S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R.W. McCarley, C. McDonald, A.M. McIntosh, S. Meier, C.J. Meijer, B. Melegh, I. Melle, R.I. Mesholam-Gately, A. Metspalu, P.T. Michie, L. Milani, V. Milanova, Y. Mokrab, D.W. Morris, O. Mors, K.C. Murphy, R.M. Murray, I. Myin-Germeys, B. Müller-Myhsok, M. Nelis, I. Nenadic, D. a. Nertney, G. Nestadt, K.K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F.A. O'Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, P. Endophenotypes International Consortium, C. Pantelis, G.N. Papadimitriou, S. Papiol, E. Parkhomenko, M.T. Pato, T. Paunio, M. Pejovic-Milovancevic, D.O. Perkins, O. Pietiläinen, J. Pimm, A.J. Pocklington, J. Powell, A. Price, A.E. Pulver, S.M. Purcell, D. Quested, H.B. Rasmussen, A. Reichenberg, M. a. Reimers, A.L. Richards, J.L. Roffman, P. Roussos, D.M. Ruderfer, V. Salomaa, A.R. Sanders, U. Schall, C.R. Schubert, T.G. Schulze, S.G. Schwab, E.M. Scolnick, R.J. Scott, L.J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J.M. Silverman, K. Sim, P. Slominsky, J.W. Smoller, H.-C. So, C. a. Spencer, E. a. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R.E. Straub, E. Strengman, J. Strohmaier, T. Scott Stroup, M. Subramaniam, J. Suvisaari, D.M. Svrakic, J.P. Szatkiewicz, E. Söderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B.T. Webb, M. Weiser, D.B. Wildenauer, N.M. Williams, S. Williams, S.H. Witt, A.R. Wolen, E.H.M. Wong, B.K. Wormley, H. Simon Xi, C.C. Zai, X. Zheng, F. Zimprich, N.R. Wray, K. Stefansson, P.M. Visscher, W. Trust Case-Control Consortium, R. Adolfsson, O. a. Andreassen, D.H.R. Blackwood, E. Bramon, J.D. Buxbaum, A.D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C.M. Hultman, N. Iwata, A. V. Jablensky, E.G. Jönsson, K.S. Kendler, G. Kirov, J. Knight, T. Lencz,

D.F. Levinson, Q.S. Li, J. Liu, A.K. Malhotra, S. a. McCarroll, A. McQuillin, J.L. Moran, P.B. Mortensen, B.J. Mowry, M.M. Nöthen, R. a. Ophoff, M.J. Owen, A. Palotie, C.N. Pato, T.L. Petryshen, D. Posthuma, M. Rietschel, B.P. Riley, D. Rujescu, P.C. Sham, P. Sklar, D. St Clair, D.R. Weinberger, J.R. Wendland, T. Werge, M.J. Daly, P.F. Sullivan, M.C. O'Donovan, Biological insights from 108 schizophrenia-associated genetic loci, Nature. 511 (2014) 421–427. doi:10.1038/nature13595.

[59] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. a R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses., Am. J. Hum. Genet. 81 (2007) 559–75. doi:10.1086/519795.

[60] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience. 4 (2015) 7. doi:10.1186/s13742-015-0047-8.

[61] J. Euesden, C.M. Lewis, P.F. O'Reilly, PRSice: Polygenic Risk Score software, Bioinformatics. 31 (2015) 1466–1468. doi:10.1093/bioinformatics/btu848.

[62] G. de Los Campos, A.I. Vazquez, R. Fernando, Y.C. Klimentidis, D. Sorensen, Prediction of complex human traits using the genomic best linear unbiased predictor., PLoS Genet. 9 (2013) e1003608. doi:10.1371/journal.pgen.1003608.

[63] G. Moser, S.H. Lee, B.J. Hayes, M.E. Goddard, N.R. Wray, P.M. Visscher, Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model, PLoS Genet. 11 (2015) 1–22. doi:10.1371/journal.pgen.1004969.

[64] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, T. Aittokallio, Regularized Machine Learning in the Genetic Prediction of Complex Traits, PLoS Genet. 10 (2014). doi:10.1371/journal.pgen.1004754.

[65] G. Abraham, A. Kowalczyk, J. Zobel, M. Inouye, Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease, Genet. Epidemiol. 37 (2013) 184–195. doi:10.1002/gepi.21698.

[66] R. Maier, G. Moser, G.-B. Chen, S. Ripke, W. Coryell, J.B. Potash, W. a. Scheftner, J. Shi, M.M. Weissman, C.M. Hultman, M. Landén, D.F. Levinson, K.S. Kendler, J.W.

Smoller, N.R. Wray, S.H. Lee, Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder, Am. J. Hum. Genet. 96 (2015) 283–294. doi:10.1016/j.ajhg.2014.12.006.

[67] J. Yang, T. Ferreira, A.P. Morris, S.E. Medland, P. a F. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.N. Weedon, R.J. Loos, T.M. Frayling, M.I. McCarthy, J.N. Hirschhorn, M.E. Goddard, P.M. Visscher, Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits, Nat. Genet. 44 (2012) 369–375. doi:10.1038/ng.2213.

[68] M.R. Robinson, A. Kleinman, M. Graff, A.A.E. Vinkhuyzen, D. Couper, M.B. Miller, W.J. Peyrot, A. Abdellaoui, B.P. Zietsch, I.M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, B.Z. Alizadeh, H.M. Boezen, L. Franke, P. van der Harst, G. Navis, M. Rots, H. Snieder, M. Swertz, B.H.R. Wolffenbuttel, C. Wijmenga, G.R. Abecasis, D. Absher, H. Alavere, E. Albrecht, H.L. Allen, P. Almgren, N. Amin, P. Amouyel, D. Anderson, A.M. Arnold, D. Arveiler, T. Aspelund, F.W. Asselbergs, T.L. Assimes, M. Atalay, A.P. Attwood, L.D. Atwood, S.J.L. Bakker, B. Balkau, A.J. Balmforth, C. Barlassina, I. Barroso, H. Basart, S. Bauer, J.S. Beckmann, J.P. Beilby, A.J. Bennett, Y. Ben-Shlomo, R.N. Bergman, S. Bergmann, S.I. Berndt, R. Biffar, A.M. Di Blasio, B.O. Boehm, M. Boehnke, H. Boeing, E. Boerwinkle, J.L. Bolton, A. Bonnefond, L.L. Bonnycastle, D.I. Boomsma, I.B. Borecki, S.R. Bornstein, N. Bouatia-Naji, G. Boucher, J.L. Bragg-Gresham, P. Brambilla, M. Bruinenberg, T.A. Buchanan, C. Buechler, G. Cadby, H. Campbell, M.J. Caulfield, C. Cavalcanti-Proença, G. Cesana, S.J. Chanock, D.I. Chasman, Y.-D.I. Chen, P.S. Chines, D.J. Clegg, L. Coin, F.S. Collins, J.M. Connell, W. Cookson, M.N. Cooper, D.C. Croteau-Chonka, L.A. Cupples, D. Cusi, F.R. Day, I.N.M. Day, G. V. Dedoussis, M. Dei, P. Deloukas, E.T. Dermitzakis, A.S. Dimas, M. Dimitriou, A.L. Dixon, M. Dörr, C.M. van Duijn, S. Ebrahim, S. Edkins, G. Eiriksdottir, K. Eisinger, N. Eklund, P. Elliott, R. Erbel, J. Erdmann, M.R. Erdos, J.G. Eriksson, T. Esko, K. Estrada, D.M. Evans, U. de Faire, T. Fall, M. Farrall, M.F. Feitosa, M.M. Ferrario, T. Ferreira, J. Ferrières, K. Fischer, E. Fisher, G. Fowkes, C.S. Fox, L. Franke, P.W. Franks, R.M. Fraser, F. Frau, T. Frayling, N.B. Freimer, P. Froguel, M. Fu, S. Gaget, A. Ganna, P. V. Gejman, D. Gentilini, E.J.C. Geus, C. Gieger, B. Gigante, A.P. Gjesing, N.L. Glazer, M.E. Goddard, A. Goel, H. Grallert, J. Gräßler, H. Grönberg, L.C. Groop, C.J. Groves, V. Gudnason, C. Guiducci, S.

Gustafsson, U. Gyllensten, A.S. Hall, P. Hall, G. Hallmans, A. Hamsten, T. Hansen, T. Haritunians, T.B. Harris, P. van der Harst, A.-L. Hartikainen, N. Hassanali, A.T. Hattersley, A.S. Havulinna, C. Hayward, N.L. Heard-Costa, A.C. Heath, J. Hebebrand, I.M. Heid, M. den Heijer, C. Hengstenberg, K.-H. Herzig, A.A. Hicks, A. Hingorani, A. Hinney, J.N. Hirschhorn, A. Hofman, C.C. Holmes, G. Homuth, J.-J. Hottenga, K.G. Hovingh, F.B. Hu, Y.-J. Hu, J.E. Huffman, J. Hui, H. Huikuri, S.E. Humphries, J. Hung, S.E. Hunt, D. Hunter, K. Hveem, E. Hyppönen, W. Igl, T. Illig, E. Ingelsson, C. Iribarren, B. Isomaa, A.U. Jackson, K.B. Jacobs, A.L. James, J.-O. Jansson, I. Jarick, M.-R. Jarvelin, K.-H. Jöckel, Å. Johansson, T. Johnson, J. Jolley, T. Jørgensen, P. Jousilahti, A. Jula, A.E. Justice, M. Kaakinen, M. Kähönen, E. Kajantie, S. Kanoni, W.H.L. Kao, L.M. Kaplan, R.C. Kaplan, J. Kaprio, K. Kapur, F. Karpe, S. Kathiresan, F. Kee, S.M. Keinanen-Kiukaanniemi, S. Ketkar, J. Kettunen, K.-T. Khaw, L.A. Kiemeney, T.O. Kilpeläinen, L. Kinnunen, M. Kivimaki, M. Kivmaki, M.M. Van der Klauw, M.E. Kleber, J.W. Knowles, W. Koenig, I. Kolcic, G. Kolovou, I.R. König, S. Koskinen, P. Kovacs, P. Kraft, A.T. Kraja, K. Kristiansson, K. Krjutškov, H.K. Kroemer, J.P. Krohn, V. Krzelj, D. Kuh, J.R. Kulzer, M. Kumari, Z. Kutalik, K. Kuulasmaa, J. Kuusisto, K. Kvaloy, M. Laakso, J.H. Laitinen, T.A. Lakka, C. Lamina, C. Langenberg, O. Lantieri, G.M. Lathrop, L.J. Launer, D.A. Lawlor, R.W. Lawrence, I.M. Leach, C. Lecoeur, S.H. Lee, T. Lehtimäki, M.F. Leitzmann, G. Lettre, D.F. Levinson, G. Li, S. Li, L. Liang, D.-Y. Lin, L. Lind, C.M. Lindgren, J. Lindström, J. Liu, A. Liuzzi, A.E. Locke, M.-L. Lokki, C. Loley, R.J.F. Loos, M. Lorentzon, J. Luan, R.N. Luben, B. Ludwig, P.A. Madden, R. Mägi, P.K.E. Magnusson, M. Mangino, P. Manunta, D. Marek, M. Marre, N.G. Martin, W. März, A. Maschio, I. Mathieson, W.L. McArdle, S.A. McCaroll, A. McCarthy, M.I. McCarthy, B. McKnight, C. Medina-Gomez, S.E. Medland, T. Meitinger, A. Metspalu, J.B.J. van Meurs, D. Meyre, K. Midthjell, E. Mihailov, L. Milani, J.L. Min, S. Moebus, M.F. Moffatt, K.L. Mohlke, C. Molony, K.L. Monda, G.W. Montgomery, V. Mooser, M.A. Morken, A.D. Morris, A.P. Morris, T.W. Mìhleisen, M. Mìller-Nurasyid, P.B. Munroe, A.W. Musk, N. Narisu, G. Navis, B.M. Neale, M. Nelis, J. Nemesh, M.J. Neville, J.S. Ngwa, G. Nicholson, M.S. Nieminen, I. Njølstad, E.A. Nohr, I.M. Nolte, K.E. North, M.M. Nöthen, D.R. Nyholt, J.R. O'Connell, C. Ohlsson, A.J. Oldehinkel, G.-J. van Ommen, K.K. Ong, B.A. Oostra, W.H. Ouwehand, C.N.A. Palmer, L.J. Palmer, A. Palotie, G. Paré, A.N. Parker, L. Paternoster, Y. Pawitan, S. Pechlivanis, J.F. Peden, N.L. Pedersen, O. Pedersen, N. Pellikka, L. Peltonen, B. Penninx, M. Perola, J.R.B. Perry, T. Person, A. Peters, M.J.

Peters, I. Pichler, K.H. Pietiläinen, C.G.P. Platou, O. Polasek, A. Pouta, C. Power, P.P. Pramstaller, M. Preuss, J.F. Price, I. Prokopenko, M.A. Province, B.M. Psaty, S. Purcell, C. Pìtter, L. Qi, T. Quertermous, A. Radhakrishnan, O. Raitakari, J.C. Randall, R. Rauramaa, N.W. Rayner, E. Rehnberg, A. Rendon, M. Ridderstråle, P.M. Ridker, S. Ripatti, A. Rissanen, F. Rivadeneira, C. Rivolta, N.R. Robertson, L.M. Rose, I. Rudan, T.E. Saaristo, H. Sager, V. Salomaa, N.J. Samani, J.G. Sambrook, A.R. Sanders, C. Sandholt, S. Sanna, J. Saramies, E.E. Schadt, A. Scherag, S. Schipf, D. Schlessinger, S. Schreiber, H. Schunkert, P.E.H. Schwarz, L.J. Scott, J. Shi, S.-Y. Shin, A.R. Shuldiner, D. Shungin, S. Signorini, K. Silander, J. Sinisalo, B. Skrobek, J.H. Smit, A.V. Smith, G.D. Smith, H. Snieder, N. Soranzo, T.I.A. Sørensen, U. Sovio, T.D. Spector, E.K. Speliotes, A. Stančáková, K. Stark, K. Stefansson, V. Steinthorsdottir, J.C. Stephens, K. Stirrups, R.P. Stolk, D.P. Strachan, R.J. Strawbridge, H.M. Stringham, M. Stumvoll, I. Surakka, A.J. Swift, A.-C. Syvanen, M.-L. Tammesoo, M. Teder-Laving, T.M. Teslovich, A. Teumer, E. V. Theodoraki, B. Thomson, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, N.J. Timpson, A. Tönjes, D.-A. Tregouet, E. Tremoli, M.D. Trip, T. Tuomi, J. Tuomilehto, J. Tyrer, M. Uda, A.G. Uitterlinden, G. Usala, M. Uusitupa, T.T. Valle, L. Vandenput, V. Vatin, S. Vedantam, F. de Vegt, S.H. Vermeulen, J. Viikari, J. Virtamo, P.M. Visscher, V. Vitart, J. V. Van Vliet-Ostaptchouk, B.F. Voight, P. Vollenweider, C.B. Volpato, H. Völzke, G. Waeber, L.L. Waite, H. Wallaschofski, G.B. Walters, Z. Wang, N.J. Wareham, R.M. Watanabe, H. Watkins, M.N. Weedon, R. Welch, R.J. Weyant, E. Wheeler, C.C. White, H.-E. Wichmann, E. Widen, S.H. Wild, G. Willemsen, C.J. Willer, T. Wilsgaard, J.F. Wilson, S. van Wingerden, B.R. Winkelmann, T.W. Winkler, D.R. Witte, J.C.M. Witteman, B.H.R. Wolffenbuttel, A. Wong, A.R. Wood, T. Workalemahu, A.F. Wright, J. Yang, J.W.G. Yarnell, L. Zgaga, J.H. Zhao, M.C. Zillikens, P. Zitting, K.T. Zondervan, S.E. Medland, N.G. Martin, P.K.E. Magnusson, W.G. Iacono, M. McGue, K.E. North, J. Yang, P.M. Visscher, Genetic evidence of assortative mating in humans, Nat. Hum. Behav. 1 (2017) 16. doi:10.1038/s41562-016-0016.

[69] B.J. Vilhjálmsson, J. Yang, H.K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E.E. Kenny, M.H. Schierup, P. De Jager, N. a. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P.M. Visscher, P. Kraft, N. Patterson, A.L. Price, Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores, Am.

J. Hum. Genet. 97 (2015) 576–592. doi:10.1016/j.ajhg.2015.09.001.

[70] N.R. Wray, S.H. Lee, D. Mehta, A. a. E. Vinkhuyzen, F. Dudbridge, C.M. Middeldorp, Research Review: Polygenic methods and their application to psychiatric traits, J. Child Psychol. Psychiatry. 55 (2014) 1068–1087. doi:10.1111/jcpp.12295.

[71] R. a Power, S. Steinberg, G. Bjornsdottir, C. a Rietveld, A. Abdellaoui, M.M. Nivard, M. Johannesson, T.E. Galesloot, J.J. Hottenga, G. Willemsen, D. Cesarini, D.J. Benjamin, P.K.E. Magnusson, F. Ullén, H. Tiemeier, A. Hofman, F.J. a van Rooij, G.B. Walters, E. Sigurdsson, T.E. Thorgeirsson, A. Ingason, A. Helgason, A. Kong, L. a Kiemeney, P. Koellinger, D.I. Boomsma, D. Gudbjartsson, H. Stefansson, K. Stefansson, Polygenic risk scores for schizophrenia and bipolar disorder predict creativity, Nat. Neurosci. 18 (2015) 953–955. doi:10.1038/nn.4040.

[72] R. a Power, K.J.H. Verweij, M. Zuhair, G.W. Montgomery, a K. Henders, a C. Heath, P. a F. Madden, S.E. Medland, N.R. Wray, N.G. Martin, Genetic predisposition to schizophrenia associated with increased use of cannabis., Mol. Psychiatry. 19 (2014) 1201–4. doi:10.1038/mp.2014.51.

[73] H.J. Jones, E. Stergiakouli, K.E. Tansey, L. Hubbard, J. Heron, M. Cannon, P. Holmans, G. Lewis, D.E.J. Linden, P.B. Jones, G. Davey Smith, M.C. O'Donovan, M.J. Owen, J.T. Walters, S. Zammit, Phenotypic Manifestation of Genetic Risk for Schizophrenia During Adolescence in the General Population, JAMA Psychiatry. 73 (2016) 221–228. doi:10.1001/jamapsychiatry.2015.3058.

[74] K.S. Kendler, The Schizophrenia Polygenic Risk Score, JAMA Psychiatry. 73 (2016) 193. doi:10.1001/jamapsychiatry.2015.2964.

[75] W.J. Peyrot, Y. Milaneschi, A. Abdellaoui, P.F. Sullivan, J.J. Hottenga, D.I. Boomsma, B.W.J.H. Penninx, Effect of polygenic risk scores on depression in childhood trauma., Br. J. Psychiatry. 205 (2014) 113–9. doi:10.1192/bjp.bp.113.143081.

[76] N. Mullins, R. a. Power, H.L. Fisher, K.B. Hanscombe, J. Euesden, R. Iniesta, D.F. Levinson, M.M. Weissman, J.B. Potash, J. Shi, R. Uher, S. Cohen-Woods, M. Rivera, L. Jones, I. Jones, N. Craddock, M.J. Owen, A. Korszun, I.W. Craig, a. E. Farmer, P. McGuffin, G. Breen, C.M. Lewis, Polygenic interactions with environmental adversity in the aetiology of major depressive disorder, Psychol. Med. 46 (2016) 759–770.

doi:10.1017/S0033291715002172.

[77] D.M. Evans, G. Davey Smith, Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality, Annu. Rev. Genomics Hum. Genet. 16 (2015) 327–350. doi:10.1146/annurev-genom-090314-050016.

[78] B.F. Voight, G.M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M.K. Jensen, G. Hindy, H. Hólm, E.L. Ding, T. Johnson, H. Schunkert, N.J. Samani, R. Clarke, J.C. Hopewell, J.F. Thompson, M. Li, G. Thorleifsson, C. Newton-Cheh, K. Musunuru, J.P. Pirruccello, D. Saleheen, L. Chen, A.F. Stewart, A. Schillert, U. Thorsteinsdottir, G. Thorgeirsson, S. Anand, J.C. Engert, T. Morgan, J. Spertus, M. Stoll, K. Berger, N. Martinelli, D. Girelli, P.P. McKeown, C.C. Patterson, S.E. Epstein, J. Devaney, M.-S. Burnett, V. Mooser, S. Ripatti, I. Surakka, M.S. Nieminen, J. Sinisalo, M.-L. Lokki, M. Perola, A. Havulinna, U. de Faire, B. Gigante, E. Ingelsson, T. Zeller, P. Wild, P.I.W. de Bakker, O.H. Klungel, A.-H. Maitland-van der Zee, B.J.M. Peters, A. de Boer, D.E. Grobbee, P.W. Kamphuisen, V.H.M. Deneer, C.C. Elbers, N.C. Onland-Moret, M.H. Hofker, C. Wijmenga, W.M. Verschuren, J.M. Boer, Y.T. van der Schouw, A. Rasheed, P. Frossard, S. Demissie, C. Willer, R. Do, J.M. Ordovas, G.R. Abecasis, M. Boehnke, K.L. Mohlke, M.J. Daly, C. Guiducci, N.P. Burtt, A. Surti, E. Gonzalez, S. Purcell, S. Gabriel, J. Marrugat, J. Peden, J. Erdmann, P. Diemert, C. Willenborg, I.R. König, M. Fischer, C. Hengstenberg, A. Ziegler, I. Buysschaert, D. Lambrechts, F. Van de Werf, K.A. Fox, N.E. El Mokhtari, D. Rubin, J. Schrezenmeir, S. Schreiber, A. Schäfer, J. Danesh, S. Blankenberg, R. Roberts, R. McPherson, H. Watkins, A.S. Hall, K. Overvad, E. Rimm, E. Boerwinkle, A. Tybjaerg-Hansen, L.A. Cupples, M.P. Reilly, O. Melander, P.M. Mannucci, D. Ardissino, D. Siscovick, R. Elosua, K. Stefansson, C.J. O'Donnell, V. Salomaa, D.J. Rader, L. Peltonen, S.M. Schwartz, D. Altshuler, S. Kathiresan, Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study, Lancet. 380 (2012) 572–580. doi:10.1016/S0140-6736(12)60312-2.

[79] C. Baigent, A. Keech, P.M. Kearney, L. Blackwell, G. Buck, C. Pollicino, A. Kirby, T. Sourjina, R. Peto, R. Collins, R. Simes, Cholesterol Treatment Trialists' (CTT) Collaborators, Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins., Lancet (London, England). 366 (2005) 1267–78. doi:10.1016/S0140-6736(05)67394-1.

[80] D. Keene, C. Price, M.J. Shun-Shin, D.P. Francis, Effect on cardiovascular risk of high density lipoprotein targeted drug treatments niacin, fibrates, and CETP inhibitors: meta-analysis of randomised controlled trials including 117,411 patients., BMJ. 349 (2014) g4379. doi:10.1136/bmj.g4379.

[81] M.J.A. Brion, K. Shakhbazov, P.M. Visscher, Calculating statistical power in Mendelian randomization studies, Int. J. Epidemiol. 42 (2013) 1497–1501. doi:10.1093/ije/dyt179.

[82] S. Burgess, A. Butterworth, S.G. Thompson, Mendelian randomization analysis with multiple genetic variants using summarized data, Genet. Epidemiol. 37 (2013) 658–665. doi:10.1002/gepi.21758.

[83] R.G. Barr, MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations, (2016). doi:https://doi.org/10.1101/078972.

[84] F.P. Hartwig, J. Bowden, C. Loret de Mola, L. Tovo-Rodrigues, G. Davey Smith, B.L. Horta, Body mass index and psychiatric disorders: a Mendelian randomization study, Sci. Rep. 6 (2016) 32730. doi:10.1038/srep32730.

[85] B.P. Prins, A. Abbasi, A. Wong, A. Vaez, I. Nolte, N. Franceschini, P.E. Stuart, J. Guterriez Achury, V. Mistry, J.P. Bradfield, A.M. Valdes, J. Bras, A. Shatunov, C. Lu, B. Han, S. Raychaudhuri, S. Bevan, M.D. Mayes, L.C. Tsoi, E. Evangelou, R.P. Nair, S.F.A. Grant, C. Polychronakos, T.R.D. Radstake, D.A. van Heel, M.L. Dunstan, N.W. Wood, A. Al-Chalabi, A. Dehghan, H. Hakonarson, H.S. Markus, J.T. Elder, J. Knight, D.E. Arking, T.D. Spector, B.P.C. Koeleman, C.M. van Duijn, J. Martin, A.P. Morris, R.K. Weersma, C. Wijmenga, P.B. Munroe, J.R.B. Perry, J.G. Pouget, Y. Jamshidi, H. Snieder, B.Z. Alizadeh, Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study, PLoS Med. 13 (2016) 1–29. doi:10.1371/journal.pmed.1001976.

[86] M. Inoshita, S. Numata, A. Tajima, M. Kinoshita, H. Umehara, M. Nakataki, M. Ikeda, S. Maruyama, H. Yamamori, T. Kanazawa, S. Shimodera, R. Hashimoto, I. Imoto, H. Yoneda, N. Iwata, T. Ohmori, A significant causal association between C-reactive protein levels and schizophrenia., Sci. Rep. 6 (2016) 26105. doi:10.1038/srep26105.

[87] S.H. Gage, M. Hickman, S. Zammit, Association between cannabis and psychosis: Epidemiologic evidence, Biol. Psychiatry. 79 (2016) 549–556. doi:10.1016/j.biopsych.2015.08.001.

[88] J. McGrath, J. Welham, J. Scott, D. Varghese, L. Degenhardt, M.R. Hayatbakhsh, R. Alati, G.M. Williams, W. Bor, J.M. Najman, Association between cannabis use and psychosis-related outcomes using sibling pair analysis in a cohort of young adults., Arch. Gen. Psychiatry. 67 (2010) 440–7. doi:10.1001/archgenpsychiatry.2010.6.

[89] S.H. Gage, H.J. Jones, S. Burgess, J. Bowden, G. Davey Smith, S. Zammit, M.R. Munafò, Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample Mendelian randomization study, Psychol. Med. 47 (2017) 971–980. doi:10.1017/S0033291716003172.

[90] J. Vaucher, B.J. Keating, A.M. Lasserre, W. Gan, D.M. Lyall, J. Ward, D.J. Smith, J.P. Pell, N. Sattar, G. Paré, M. V Holmes, Cannabis use and risk of schizophrenia: a Mendelian randomization study, Mol. Psychiatry. (2017) 1–6. doi:10.1038/mp.2016.252.

[91] A.E. Taylor, S. Burgess, J.J. Ware, S.H. Gage, J.B. Richards, G. Davey Smith, M.R. Munafò, Investigating causality in the association between 25(OH)D and schizophrenia, Sci. Rep. 6 (2016) 26496. doi:10.1038/srep26496.

[92] J.H. Bjørngaard, D. Gunnell, M.B. Elvestad, G. Davey Smith, F. Skorpen, H. Krokan, L. Vatten, P. Romundstad, The causal role of smoking in anxiety and depression: a Mendelian randomization analysis of the HUNT study., Psychol. Med. 43 (2013) 711–9. doi:10.1017/S0033291712001274.

[93] S.H. Gage, G.D. Smith, S. Zammit, M. Hickman, M.R. Munafò, Using Mendelian randomisation to infer causality in depression and anxiety research., Depress. Anxiety. 30 (2013) 1185–93. doi:10.1002/da.22150.

[94] A. Sekar, A.R. Bialas, H. de Rivera, A. Davis, T.R. Hammond, N. Kamitaki, K. Tooley, J. Presumey, M. Baum, V. Van Doren, G. Genovese, S.A. Rose, R.E. Handsaker, Schizophrenia Working Group of the Psychiatric Genomics Consortium, M.J. Daly, M.C. Carroll, B. Stevens, S.A. McCarroll, Schizophrenia risk from complex variation of complement component 4., Nature. 530 (2016) 177–83. doi:10.1038/nature16549.

[95] M. Claussnitzer, S.N. Dankel, K.-H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I.S. Sousa, J.L. Beaudry, V. Puviindran, N.A. Abdennur, J. Liu, P.-A. Svensson, Y.-H. Hsu, D.J. Drucker, G. Mellgren, C.-C. Hui, H. Hauner, M. Kellis, FTO Obesity Variant Circuitry and Adipocyte Browning in Humans., N. Engl. J. Med. 373 (2015) 895–907. doi:10.1056/NEJMoa1502214.

[96] K.K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W.J. Housley, S. Beik, N. Shoresh, H. Whitton, R.J.H. Ryan, A.A. Shishkin, M. Hatan, M.J. Carrasco-Alfonso, D. Mayer, C.J. Luckey, N.A. Patsopoulos, P.L. De Jager, V.K. Kuchroo, C.B. Epstein, M.J. Daly, D.A. Hafler, B.E. Bernstein, Genetic and epigenetic fine mapping of causal autoimmune disease variants., Nature. 518 (2015) 337–43. doi:10.1038/nature13835.

[97] J.K. Pickrell, T. Berisa, J.Z. Liu, L. Ségurel, J.Y. Tung, D.A. Hinds, L. Segurel, J.Y. Tung, D.A. Hinds, Detection and interpretation of shared genetic influences on 42 human traits, Nat. Genet. 48 (2016) 709–717. doi:10.1038/ng.3570.

[98] B.M. Neale, P.C. Sham, The future of association studies: gene-based analysis and replication., Am. J. Hum. Genet. 75 (2004) 353–62. doi:10.1086/423901.

[99] E.R. Gamazon, H.E. Wheeler, K.P. Shah, S. V Mozaffari, K. Aquino-Michaels, R.J. Carroll, A.E. Eyler, J.C. Denny, D.L. Nicolae, N.J. Cox, H.K. Im, A gene-based association method for mapping traits using reference transcriptome data, Nat. Genet. 47 (2015) 1091–1098. doi:10.1038/ng.3367.

[100] T.J. Hohman, L. Dumitrescu, N.J. Cox, A.L. Jefferson, Genetic resilience to amyloid related cognitive decline, Brain Imaging Behav. 11 (2017) 401–409. doi:10.1007/s11682-016-9615-5.

[101] M.A.R. Ferreira, R. Jansen, G. Willemsen, B. Penninx, L.M. Bain, C.T. Vicente, J.A. Revez, M.C. Matheson, J. Hui, J.Y. Tung, S. Baltic, P. Le Souëf, G.W. Montgomery, N.G. Martin, C.F. Robertson, A. James, P.J. Thompson, D.I. Boomsma, J.L. Hopper, D.A. Hinds, R.B. Werder, S. Phipps, Gene-based analysis of regulatory variants identifies 4 putative novel asthma risk genes related to nucleotide synthesis and signaling, J. Allergy Clin. Immunol. 139 (2017) 1148–1157. doi:10.1016/j.jaci.2016.07.017.

[102] A. Barbeira, K.P. Shah, J.M. Torres, H.E. Wheeler, E.S. Torstenson, T. Edwards, T.

Garcia, G.I. Bell, D. Nicolae, N.J. Cox, H.K. Im, MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results, bioRxiv. (2016) 45260. doi:10.1101/045260.

[103] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B.W.J.H. Penninx, R. Jansen, E.J.C. de Geus, D.I. Boomsma, F.A. Wright, P.F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A.J. Lusis, T. Lehtimäki, E. Raitoharju, M. Kähönen, I. Seppälä, O.T. Raitakari, J. Kuusisto, M. Laakso, A.L. Price, P. Pajukanta, B. Pasaniuc, Integrative approaches for large-scale transcriptome-wide association studies., Nat. Genet. 48 (2016) 245–52. doi:10.1038/ng.3506.

[104] N. Mancuso, H. Shi, P. Goddard, G. Kichaev, A. Gusev, B. Pasaniuc, Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits, Am. J. Hum. Genet. 100 (2017) 473–487. doi:10.1016/j.ajhg.2017.01.031.

[105] T.H. Pers, J.M. Karjalainen, Y. Chan, H.-J. Westra, A.R. Wood, J. Yang, J.C. Lui, S. Vedantam, S. Gustafsson, T. Esko, T. Frayling, E.K. Speliotes, M. Boehnke, S. Raychaudhuri, R.S.N. Fehrmann, J.N. Hirschhorn, L. Franke, Biological interpretation of genome-wide association studies using predicted gene functions, Nat. Commun. 6 (2015) 5890. doi:10.1038/ncomms6890.

[106] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M.R. Robinson, J.E. Powell, G.W. Montgomery, M.E. Goddard, N.R. Wray, P.M. Visscher, J. Yang, Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, Nat. Genet. 48 (2016) 481–487. doi:10.1038/ng.3538.

[107] K. Shah, H.E. Wheeler, E.R. Gamazon, D.L. Nicolae, N.J. Cox, H.K. Im, Genetic predictors of gene expression associated with risk of bipolar disorder, 2016. doi:10.1101/043752.

[108] A. Gusev, N. Mancuso, H.K. Finucane, Y. Reshef, L. Song, A. Safi, E. Oh, S. McCaroll, B. Neale, R. Ophoff, M.C. O'Donovan, N. Katsanis, G.E. Crawford, P.F. Sullivan, B. Pasaniuc, A.L. Price, Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights, bioRxiv. (2016) 67355. doi:10.1101/067355.

[109] J. Flint, K.S. Kendler, The genetics of major depression., Neuron. 81 (2014) 484–503. doi:10.1016/j.neuron.2014.01.027.

[110] S.S. Jeste, D.H. Geschwind, Disentangling the heterogeneity of autism spectrum disorder through genetic findings., Nat. Rev. Neurol. 10 (2014) 74–81. doi:10.1038/nrneurol.2013.278.

[111] S. Kapur, A.G. Phillips, T.R. Insel, Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?, Mol. Psychiatry. 17 (2012) 1174–9. doi:10.1038/mp.2012.105.

[112] B. Han, J.G. Pouget, K. Slowikowski, E. Stahl, C.H. Lee, D. Diogo, X. Hu, Y.R. Park, E. Kim, P.K. Gregersen, S.R. Dahlqvist, J. Worthington, J. Martin, S. Eyre, L. Klareskog, T. Huizinga, W.-M. Chen, S. Onengut-Gumuscu, S.S. Rich, M.D.D.W.G. of the P.G. Consortium, N.R. Wray, S. Raychaudhuri, A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases, Nat Genet. 48 (2016) 803–810. doi:10.1038/ng.3572.

[113] J. Liley, J.A. Todd, C. Wallace, A method for identifying genetic heterogeneity within phenotypically defined disease subgroups, Nat. Genet. 49 (2016) 310–316. doi:10.1038/ng.3751.

[114] S.J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D.J. Hunter, G. Thomas, J.N. Hirschhorn, G. Abecasis, D. Altshuler, J.E. Bailey-Wilson, L.D. Brooks, L.R. Cardon, M. Daly, P. Donnelly, J.F. Fraumeni, N.B. Freimer, D.S. Gerhard, C. Gunter, A.E. Guttmacher, M.S. Guyer, E.L. Harris, J. Hoh, R. Hoover, C.A. Kong, K.R. Merikangas, C.C. Morton, L.J. Palmer, E.G. Phimister, J.P. Rice, J. Roberts, C. Rotimi, M.A. Tucker, K.J. Vogan, S. Wacholder, E.M. Wijsman, D.M. Winn, F.S. Collins, Replicating genotype-phenotype associations., Nature. 447 (2007) 655–60. doi:10.1038/447655a.

[115] D.S. Falconer, The inheritance of liability to certain diseases, estimated from the incidence among relatives, Ann. Hum. Genet. 29 (1965) 51–76. doi:10.1111/j.1469-1809.1965.tb00500.x.

[116] T. Reich, J.W. James, C.A. Morris, The use of multiple thresholds in determining the

mode of transmission of semi-continuous traits., Ann. Hum. Genet. 36 (1972) 163–84. doi:10.1111/j.1469-1809.1972.tb00767.x.

[117] E.R. Dempster, I.M. Lerner, Heritability of Threshold Characters., Genetics. 35 (1950) 212–36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1209482&tool=pmcentrez&rendertype=abstract.

[118] A. Tenesa, C.S. Haley, The heritability of human disease: estimation, uses and abuses., Nat. Rev. Genet. 14 (2013) 139–149. doi:10.1038/nrg3377.

[119] N.J. Risch, Searching for genetic determinants in the new millennium, Nature. 405 (2000) 847–856. doi:10.1038/35015718.

[120] M.J. Bamshad, S.B. Ng, A.W. Bigham, H.K. Tabor, M.J. Emond, D.A. Nickerson, J. Shendure, Exome sequencing as a tool for Mendelian disease gene discovery, Nat. Rev. Genet. 12 (2011) 745–755. doi:10.1038/nrg3031.

[121] E.T. Cirulli, D.B. Goldstein, Uncovering the roles of rare variants in common disease through whole-genome sequencing., Nat. Rev. Genet. 11 (2010) 415–425. doi:10.1038/nrg2779.

[122] L.E. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, B.W. van Bon, a Hoischen, B.B. de Vries, H.G. Brunner, J. a Veltman, A de novo paradigm for mental retardation, Nat Genet. 42 (2010) 1109–1112. doi:10.1038/ng.712.

[123] S.J. Sanders, M.T. Murtha, A.R. Gupta, J.D. Murdoch, M.J. Raubeson, A.J. Willsey, A.G. Ercan-Sencicek, N.M. DiLullo, N.N. Parikshak, J.L. Stein, M.F. Walker, G.T. Ober, N.A. Teran, Y. Song, P. El-Fishawy, R.C. Murtha, M. Choi, J.D. Overton, R.D. Bjornson, N.J. Carriero, K.A. Meyer, K. Bilguvar, S.M. Mane, N. Šestan, R.P. Lifton, M. Günel, K. Roeder, D.H. Geschwind, B. Devlin, M.W. State, De novo mutations revealed by whole-exome sequencing are strongly associated with autism, Nature. 485 (2012) 237–241. doi:10.1038/nature10945.

[124] M. Fromer, A.J. Pocklington, D.H. Kavanagh, H.J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D.M. Ruderfer, N. Carrera, I. Humphreys, J.S. Johnson, P. Roussos, D.D. Barker, E. Banks, V. Milanova, S.G. Grant, E. Hannon, S.A. Rose,

K. Chambert, M. Mahajan, E.M. Scolnick, J.L. Moran, G. Kirov, A. Palotie, S.A. McCarroll, P. Holmans, P. Sklar, M.J. Owen, S.M. Purcell, M.C. O'Donovan, De novo mutations in schizophrenia implicate synaptic networks, Nature. 506 (2014) 179–184. doi:10.1038/nature12929.

[125] A. Kiezun, K. Garimella, R. Do, N.O. Stitziel, M. Benjamin, P.J. Mclaren, N. Gupta, P. Sklar, P.F. Sullivan, J.L. Moran, C.M. Hultman, P. Lichtenstein, P. Magnusson, Exome sequencing and the genetic basis of complex traits, NIH Public Access. 44 (2013) 623–630. doi:10.1038/ng.2303.Exome.

[126] L. Vadlamudi, L.M. Dibbens, K.M. Lawrence, X. Iona, J.M. McMahon, W. Murrell, A. Mackay-Sim, I.E. Scheffer, S.F. Berkovic, Timing of De Novo Mutagenesis — A Twin Study of Sodium-Channel Mutations, N. Engl. J. Med. 363 (2010) 1335–1340. doi:10.1056/NEJMoa0910752.

[127] J. Gratten, P.M. Visscher, B.J. Mowry, N.R. Wray, Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease., Nat. Genet. 45 (2013) 234–8. doi:10.1038/ng.2555.

[128] J. Yang, P.M. Visscher, N.R. Wray, Sporadic cases are the norm for complex disease., Eur. J. Hum. Genet. 18 (2010) 1039–1043. doi:10.1038/ejhg.2009.177.

[129] C. Smith, Heritability of liability and concordance in monozygous twins, Ann. Hum. Genet. 34 (1970) 85–91. doi:10.1111/j.1469-1809.1970.tb00223.x.

[130] J.S. Witte, P.M. Visscher, N.R. Wray, The contribution of genetic variants to disease depends on the ruler, Nat. Rev. Genet. 15 (2014) 765–776. doi:10.1038/nrg3786.

[131] I.J. Deary, I.J. Deary, J. Yang, J. Yang, G. Davies, G. Davies, S.E. Harris, S.E. Harris, A. Tenesa, A. Tenesa, D. Liewald, D. Liewald, M. Luciano, M. Luciano, L.M. Lopez, L.M. Lopez, A.J. Gow, A.J. Gow, J. Corley, J. Corley, P. Redmond, P. Redmond, H.C. Fox, H.C. Fox, S.J. Rowe, S.J. Rowe, P. Haggarty, P. Haggarty, G. McNeill, G. McNeill, M.E. Goddard, M.E. Goddard, D.J. Porteous, D.J. Porteous, L.J. Whalley, L.J. Whalley, J.M. Starr, J.M. Starr, P.M. Visscher, P.M. Visscher, Genetic contributions to stability and change in intelligence from childhood to old age., Nature. 482 (2012) 212–215. doi:10.1038/nature10781.

[132] J. Yang, T.A. Manolio, L.R. Pasquale, E. Boerwinkle, N. Caporaso, J.M. Cunningham,

M. de Andrade, B. Feenstra, E. Feingold, M.G. Hayes, W.G. Hill, M.T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R.A. Mathias, M. Melbye, E. Pugh, M.C. Cornelis, B.S. Weir, M.E. Goddard, P.M. Visscher, Genome partitioning of genetic variation for complex traits using common SNPs, Nat. Genet. 43 (2011) 519–525. doi:10.1038/ng.823.

[133] H.C. So, M. Li, P.C. Sham, Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study, Genet. Epidemiol. 35 (2011) 447–456. doi:10.1002/gepi.20593.

[134] J.E. Powell, P.M. Visscher, M.E. Goddard, Reconciling the analysis of IBD and IBS in complex trait studies., Nat. Rev. Genet. 11 (2010) 800–5. doi:10.1038/nrg2865.

[135] P.M. Visscher, W.G. Hill, N.R. Wray, Heritability in the genomics era--concepts and misconceptions., Nat. Rev. Genet. 9 (2008) 255–66. doi:10.1038/nrg2322.

[136] O. Zuk, E. Hechter, S.R. Sunyaev, E.S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 1193–8. doi:10.1073/pnas.1119675109.

[137] S. Stringer, E.M. Derks, R.S. Kahn, W.G. Hill, N.R. Wray, Assumptions and properties of limiting pathway models for analysis of epistasis in complex traits., PLoS One. 8 (2013) e68913. doi:10.1371/journal.pone.0068913.

[138] W.G. Hill, M.E. Goddard, P.M. Visscher, Data and theory point to mainly additive genetic variance for complex traits, PLoS Genet. 4 (2008). doi:10.1371/journal.pgen.1000008.

[139] T.F.C. Mackay, Epistasis and quantitative traits: using model organisms to study gene-gene interactions., Nat. Rev. Genet. 15 (2013) 22–33. doi:10.1038/nrg3627.

[140] E.C. Dunn, M. Uddin, S. V. Subramanian, J.W. Smoller, S. Galea, K.C. Koenen, Research Review: Gene-environment interaction research in youth depression - A systematic review with recommendations for future research, J. Child Psychol. Psychiatry Allied Discip. 52 (2011) 1223–1238. doi:10.1111/j.1469-7610.2011.02466.x.

[141] L.E. Duncan, M.C. Keller, A critical review of the first 10 years of candidate gene-by-

environment interaction research in psychiatry., Am. J. Psychiatry. 168 (2011) 1041–9. doi:10.1176/appi.ajp.2011.11020191.

[142] A. Gusev, G. Bhatia, N. Zaitlen, B.J. Vilhjalmsson, D. Diogo, E.A. Stahl, P.K. Gregersen, J. Worthington, L. Klareskog, S. Raychaudhuri, R.M. Plenge, B. Pasaniuc, A.L. Price, Quantifying Missing Heritability at Known GWAS Loci, PLoS Genet. 9 (2013) 10–14. doi:10.1371/journal.pgen.1003993.

[143] S. Ripke, C. O'Dushlaine, K. Chambert, J.L. Moran, A.K. Kähler, S. Akterin, S.E. Bergen, A.L. Collins, J.J. Crowley, M. Fromer, Y. Kim, S.H. Lee, P.K.E. Magnusson, N. Sanchez, E.A. Stahl, S. Williams, N.R. Wray, K. Xia, F. Bettella, A.D. Borglum, B.K. Bulik-Sullivan, P. Cormican, N. Craddock, C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M.L. Hamshere, P. Holmans, D.M. Hougaard, K.S. Kendler, K. Lin, D.W. Morris, O. Mors, P.B. Mortensen, B.M. Neale, F.A. O'Neill, M.J. Owen, M.P. Milovancevic, D. Posthuma, J. Powell, A.L. Richards, B.P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A.B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari, S. Tosato, M. Verhage, J.T. Walters, D.F. Levinson, P. V Gejman, C. Laurent, B.J. Mowry, M.C. O'Donovan, A.E. Pulver, S.G. Schwab, D.B. Wildenauer, F. Dudbridge, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F.B. Lerer, K.-Y. Liang, W. Maier, J. Mallet, D.A. Nertney, G. Nestadt, N. Norton, G.N. Papadimitriou, R. Ribble, A.R. Sanders, J.M. Silverman, D. Walsh, N.M. Williams, B. Wormley, M.J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, J. Hall, C. Iyegbe, A. Jablensky, R.S. Kahn, L. Kalaydjieva, S. Lawrie, C.M. Lewis, D.H. Linszen, I. Mata, A. McIntosh, R.M. Murray, R.A. Ophoff, J. Van Os, M. Walshe, M. Weisbrod, D. Wiersma, P. Donnelly, I. Barroso, J.M. Blackwell, M.A. Brown, J.P. Casas, A.P. Corvin, P. Deloukas, A. Duncanson, J. Jankowski, H.S. Markus, C.G. Mathew, C.N.A. Palmer, R. Plomin, A. Rautanen, S.J. Sawcer, R.C. Trembath, A.C. Viswanathan, N.W. Wood, C.C.A. Spencer, G. Band, C. Bellenguez, C. Freeman, G. Hellenthal, E. Giannoulatou, M. Pirinen, R.D. Pearson, A. Strange, Z. Su, D. Vukcevic, C. Langford, S.E. Hunt, S. Edkins, R. Gwilliam, H. Blackburn, S.J. Bumpstead, S. Dronov, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O.T. McCann, J. Liddle, S.C. Potter, R. Ravindrarajah, M. Ricketts, A. Tashakkori-Ghanbaria, M.J. Waller, P. Weston, S. Widaa, P. Whittaker, M.I. McCarthy, K. Stefansson, E. Scolnick, S. Purcell, S.A. McCarroll, P. Sklar, C.M. Hultman, P.F. Sullivan, Genome-wide association analysis

identifies 13 new risk loci for schizophrenia., Nat. Genet. 45 (2013) 1150–9. doi:10.1038/ng.2742.

[144] A.J. Hannan, TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease., Discov. Med. 10 (2010) 314–21. http://www.ncbi.nlm.nih.gov/pubmed/21034672 (accessed April 10, 2017).

[145] A.J. Hannan, Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for “missing heritability,” Trends Genet. 26 (2010) 59–65. doi:10.1016/j.tig.2009.11.008.

[146] N.R. Wray, I.I. Gottesman, Using summary data from the Danish National Registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder, Front. Genet. 3 (2012) 1–12. doi:10.3389/fgene.2012.00118.

[147] B. Han, D. Diogo, S. Eyre, H. Kallberg, A. Zhernakova, J. Bowes, L. Padyukov, Y. Okada, M.A. González-Gay, S. Rantapää-Dahlqvist, J. Martin, T.W.J. Huizinga, R.M. Plenge, J. Worthington, P.K. Gregersen, L. Klareskog, P.I.W. De Bakker, S. Raychaudhuri, Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity, Am. J. Hum. Genet. 94 (2014) 522–532. doi:10.1016/j.ajhg.2014.02.013.

[148] Y. Wang, J.G.M. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-Van Gelder, J. Yu, T. Jatkoe, E.M.J.J. Berns, D. Atkins, J.A. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet. 365 (2005) 671–679. doi:10.1016/S0140-6736(05)17947-1.

[149] D.J. Slamon, G.M. Clark, S.G. Wong, W.J. Levin, A. Ullrich, W.L. McGuire, Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene., Science. 235 (1987) 177–82. http://www.ncbi.nlm.nih.gov/pubmed/3798106 (accessed April 10, 2017).

[150] R. Ferraldeschi, W.G. Newman, Pharmacogenetics and pharmacogenomics: a clinical reality., Ann. Clin. Biochem. 48 (2011) 410–417. doi:10.1258/acb.2011.011084.

[151] G.B. Chen, S.H. Lee, M.J.A. Brion, G.W. Montgomery, N.R. Wray, G.L. Radford-Smith, P.M. Visscher, Estimation and partitioning of (co)heritability of inflammatory

bowel disease from GWAS and immunochip data, Hum. Mol. Genet. 23 (2014) 4710–4720. doi:10.1093/hmg/ddu174.

[152] L. Jostins, S. Ripke, R.K. Weersma, R.H. Duerr, D.P. McGovern, K.Y. Hui, J.C. Lee, L.P. Schumm, Y. Sharma, C. a Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S.L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A.N. Ananthakrishnan, V. Andersen, J.M. Andrews, L. Baidoo, T. Balschun, P. a Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K.L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L.R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T.H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I.C. Lawrance, C.W. Lees, E. Louis, G. Mahy, J. Mansfield, A.R. Morgan, C. Mowat, W. Newman, O. Palmieri, C.Y. Ponsioen, U. Potocnik, N.J. Prescott, M. Regueiro, J.I. Rotter, R.K. Russell, J.D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. a Simms, J. Sventoraityte, S.R. Targan, K.D. Taylor, M. Tremelling, H.W. Verspaget, M. De Vos, C. Wijmenga, D.C. Wilson, J. Winkelmann, R.J. Xavier, S. Zeissig, B. Zhang, C.K. Zhang, H. Zhao, M.S. Silverberg, V. Annese, H. Hakonarson, S.R. Brant, G. Radford-Smith, C.G. Mathew, J.D. Rioux, E.E. Schadt, M.J. Daly, A. Franke, M. Parkes, S. Vermeire, J.C. Barrett, J.H. Cho, Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease., Nature. 491 (2012) 119–24. doi:10.1038/nature11582.

[153] S. Purcell, S.S. Cherny, P.C. Sham, Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits., Bioinformatics. 19 (2003) 149–150. doi:10.1093/bioinformatics/19.1.149.

[154] A.S. Wiik, W.J. van Venrooij, G.J.M. Pruijn, All you wanted to know about anti-CCP but were afraid to ask., Autoimmun. Rev. 10 (2010) 90–3. doi:10.1016/j.autrev.2010.08.009.

[155] P. Gibson, Y. Tong, G. Robinson, M.C. Thompson, D.S. Currle, C. Eden, T.A. Kranenburg, T. Hogg, H. Poppleton, J. Martin, D. Finkelstein, S. Pounds, A. Weiss, Z. Patay, M. Scoggins, R. Ogg, Y. Pei, Z.-J. Yang, S. Brun, Y. Lee, F. Zindy, J.C. Lindsey, M.M. Taketo, F.A. Boop, R.A. Sanford, A. Gajjar, S.C. Clifford, M.F. Roussel, P.J. McKinnon, D.H. Gutmann, D.W. Ellison, R. Wechsler-Reya, R.J. Gilbertson, Subtypes of medulloblastoma have distinct developmental origins., Nature. 468 (2010) 1095–9.

doi:10.1038/nature09587.

[156] M.R. Trusheim, E.R. Berndt, F.L. Douglas, Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers., Nat. Rev. Drug Discov. 6 (2007) 287–93. doi:10.1038/nrd2251.

[157] N.R. Wray, R. Maier, Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability, Curr. Epidemiol. Reports. 1 (2014) 220–227. doi:10.1007/s40471-014-0023-3.

[158] B. Han, J.G. Pouget, K. Slowikowski, E. Stahl, C.H. Lee, D. Diogo, X. Hu, Y.R. Park, E. Kim, P.K. Gregersen, S.R. Dahqvist, J. Worthington, S. Eyre, L. Klareskog, T. Huizinga, W.-M. Chen, S. Onengut-Gumuscu, S.S. Rich, N. Wray, S. Raychaudhuri, Using genotype data to distinguish pleiotropy from heterogeneity: deciphering coheritability in autoimmune and neuropsychiatric diseases, Cold Spring Harbor Labs Journals, 2015. doi:10.1101/030783.

[159] J. Arnedo, D.M. Svrakic, C. Del Val, R. Romero-Zaliz, H. Hernández-Cuervo, A.H. Fanous, M.T. Pato, C.N. Pato, G.A. de Erausquin, C.R. Cloninger, I. Zwir, Uncovering the Hidden Risk Architecture of the Schizophrenias: Confirmation in Three Independent Genome-Wide Association Studies., Am. J. Psychiatry. (2014) 1–15. doi:10.1176/appi.ajp.2014.14040435.

[160] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika. 29 (1964) 1–27. doi:10.1007/BF02289565.

[161] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 9362–7. doi:10.1073/pnas.0903103106.

[162] S.M. Purcell, J.L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S.E. Bergen, A. Kähler, L. Duncan, E. Stahl, G. Genovese, E. Fernández, M.O. Collins, N.H. Komiyama, J.S. Choudhary, P.K.E. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S.G.N. Grant, S.J. Haggarty, S. Gabriel, E.M. Scolnick, E.S. Lander, C.M. Hultman, P.F. Sullivan, S. a McCarroll, P. Sklar, A polygenic burden of rare disruptive mutations in schizophrenia., Nature.

506 (2014) 185–90. doi:10.1038/nature12975.

[163] X. Zhou, P. Carbonetto, M. Stephens, Polygenic modeling with bayesian sparse linear mixed models., PLoS Genet. 9 (2013) e1003264. doi:10.1371/journal.pgen.1003264.

[164] D. Speed, D.J. Balding, MultiBLUP: improved SNP-based prediction for complex traits., Genome Res. 24 (2014) 1550–7. doi:10.1101/gr.169375.113.

[165] S.H. Lee, J.H.J. van der Werf, B.J. Hayes, M.E. Goddard, P.M. Visscher, Predicting unobserved phenotypes for complex traits from whole-genome SNP data., PLoS Genet. 4 (2008) e1000231. doi:10.1371/journal.pgen.1000231.

[166] Z. Zhang, U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, J. Li, H. Simianer, Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies, PLoS One. 9 (2014) 1–12. doi:10.1371/journal.pone.0093017.

[167] Z. Wei, K. Wang, H.-Q. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J.T. Glessner, R. Chiavacci, C. Stanley, D. Monos, S.F.A. Grant, C. Polychronakos, H. Hakonarson, From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes., PLoS Genet. 5 (2009) e1000678. doi:10.1371/journal.pgen.1000678.

[168] C. Li, C. Yang, J. Gelernter, H. Zhao, Improving genetic risk prediction by leveraging pleiotropy., Hum. Genet. 133 (2014) 639–50. doi:10.1007/s00439-013-1401-5.

[169] H.D. Daetwyler, B. Villanueva, J.A. Woolliams, Accuracy of predicting the genetic risk of disease using a genome-wide approach., PLoS One. 3 (2008) e3395. doi:10.1371/journal.pone.0003395.

[170] C.R. Henderson, R.L. Quaas, Multiple trait evaluation using relatives records, J. Anim. Sci. 43 (1976) 1188–1197. doi:10.2527/jas1976.4361188x.

[171] G. Guo, F. Zhao, Y. Wang, Y. Zhang, L. Du, G. Su, Comparison of single-trait and multiple-trait genomic prediction models., BMC Genet. 15 (2014) 30. doi:10.1186/1471-2156-15-30.

[172] G. Project, E. Asia, S. Africa, S. Figs, S. Tables, An integrated map of genetic variation

from 1,092 human genomes, Nature. 135 (2012) 0–9. doi:10.1038/nature11632.

[173] D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, P.E. Bonnen, P.I.W. de Bakker, P. Deloukas, S.B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L.R. Lewis, Y. Ren, D. Wheeler, D.M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J.M. Korn, K. Kristiansson, C. Lee, S.A. McCarrol, J. Nemesh, A. Keinan, S.B. Montgomery, S. Pollack, A.L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M.J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q. Zhang, M.J.R. Ghori, R. McGinnis, W. McLaren, F. Takeuchi, S.R. Grossman, I. Shlyakhter, E.B. Hostetter, P.C. Sabeti, C.A. Adebamowo, M.W. Foster, D.R. Gordon, J. Licinio, M.C. Manca, P.A. Marshall, I. Matsuda, D. Ngare, V.O. Wang, D. Reddy, C.N. Rotimi, C.D. Royal, R.R. Sharp, C. Zeng, L.D. Brooks, J.E. McEwen, Integrating common and rare genetic variation in diverse human populations., Nature. 467 (2010) 52–8. doi:10.1038/nature09298.

[174] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D.S. Pine, K. Quinn, C. Sanislow, P. Wang, Research domain criteria (RDoC): toward a new classification framework for research on mental disorders., Am. J. Psychiatry. 167 (2010) 748–51. doi:10.1176/appi.ajp.2010.09091379.

[175] S.H. Lee, J.H.J. van der Werf, An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree., Genet. Sel. Evol. 38 25–43. doi:10.1051/gse:2005025.

[176] S.H. Lee, J. Yang, M.E. Goddard, P.M. Visscher, N.R. Wray, Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood., Bioinformatics. 28 (2012) 2540–2. doi:10.1093/bioinformatics/bts474.

[177] B.J. Hayes, P.M. Visscher, M.E. Goddard, Increased accuracy of artificial selection by using the realized relationship matrix., Genet. Res. (Camb). 91 (2009) 47–60. doi:10.1017/S0016672308009981.

[178] P.M. VanRaden, Efficient methods to compute genomic predictions., J. Dairy Sci. 91 (2008) 4414–23. doi:10.3168/jds.2007-0980.

[179] I. Strandén, D.J. Garrick, Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit., J. Dairy Sci. 92 (2009) 2971–5. doi:10.3168/jds.2008-1929.

[180] S. Raychaudhuri, J.M. Korn, S.A. McCarroll, D. Altshuler, P. Sklar, S. Purcell, M.J. Daly, Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function., PLoS Genet. 6 (2010) e1001097. doi:10.1371/journal.pgen.1001097.

[181] C. Group, P.G. Consortium, Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis., Lancet. 381 (2013) 1371–9. doi:10.1016/S0140-6736(12)62129-1.

[182] O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E.-M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, Development of a Virtual Screening Method for Identification of “Frequent Hitters” in Compound Libraries, J. Med. Chem. 45 (2002) 137–142. doi:10.1021/jm010934d.

[183] S.E. Bergen, C.T. O’Dushlaine, S. Ripke, P.H. Lee, D.M. Ruderfer, S. Akterin, J.L. Moran, K.D. Chambert, R.E. Handsaker, L. Backlund, U. Ösby, S. McCarroll, M. Landen, E.M. Scolnick, P.K.E. Magnusson, P. Lichtenstein, C.M. Hultman, S.M. Purcell, P. Sklar, P.F. Sullivan, Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder., Mol. Psychiatry. 17 (2012) 880–6. doi:10.1038/mp.2012.73.

[184] J. Shi, J.B. Potash, J.A. Knowles, M.M. Weissman, W. Coryell, W.A. Scheftner, W.B. Lawson, J.R. DePaulo, P. V Gejman, A.R. Sanders, J.K. Johnson, P. Adams, S. Chaudhury, D. Jancic, O. Evgrafov, A. Zvinyatskovskiy, N. Ertman, M. Gladis, K. Neimanas, M. Goodell, N. Hale, N. Ney, R. Verma, D. Mirel, P. Holmans, D.F. Levinson, Genome-wide association study of recurrent early-onset major depressive disorder., Mol. Psychiatry. 16 (2011) 193–201. doi:10.1038/mp.2009.124.

[185] M.R. Nelson, K. Bryc, K.S. King, A. Indap, A.R. Boyko, J. Novembre, L.P. Briley, Y. Maruyama, D.M. Waterworth, G. Waeber, P. Vollenweider, J.R. Oksenberg, S.L.

Hauser, H.A. Stirnadel, J.S. Kooner, J.C. Chambers, B. Jones, V. Mooser, C.D. Bustamante, A.D. Roses, D.K. Burns, M.G. Ehm, E.H. Lai, The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research., Am. J. Hum. Genet. 83 (2008) 347–58. doi:10.1016/j.ajhg.2008.08.005.

[186] C.-Y. Chen, S. Pollack, D.J. Hunter, J.N. Hirschhorn, P. Kraft, A.L. Price, Improved ancestry inference using weights from external reference panels., Bioinformatics. 29 (2013) 1399–406. doi:10.1093/bioinformatics/btt144.

[187] M. Kirkpatrick, D. Lofsvold, M. Bulmer, Analysis of the inheritance, selection and evolution of growth trajectories., Genetics. 124 (1990) 979–993. http://www.genetics.org/content/124/4/979 (accessed January 14, 2016).

[188] K. Meyer, W.G. Hill, Estimation of genetic and phenotypic covariance functions for longitudinal or “repeated” records by restricted maximum likelihood, Livest. Prod. Sci. 47 (1997) 185–200. doi:10.1016/S0301-6226(96)01414-5.

[189] S.H. Lee, N.R. Wray, Novel genetic analysis for case-control genome-wide association studies: quantification of power and genomic prediction accuracy., PLoS One. 8 (2013) e71494. doi:10.1371/journal.pone.0071494.

[190] A. Korte, B.J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, M. Nordborg, A mixed-model approach for genome-wide association studies of correlated traits in structured populations., Nat. Genet. 44 (2012) 1066–71. doi:10.1038/ng.2376.

[191] A. Gilmour, R. Thompson, B. Cullis, Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models, Biometrics. 51 (1995). http://www.research.ed.ac.uk/portal/en/publications/average-information-reml-an-efficient-algorithm-for-variance-parameter-estimation-in-linear-mixed-models(d489eab3-92de-4c44-a99b-0042ed11e49c)/export.html (accessed January 14, 2016).

[192] M. Lynch, B. Walsh, others, Genetics and analysis of quantitative traits, 1st editio, Sinauer Sunderland, MA, 1998.

[193] S.R. (Shayle R.. Searle, G. Casella, C.E. McCulloch, Variance components, Wiley, 2006.

[194] D.L. Johnson, R. Thompson, Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information, J. Dairy Sci. 78 (1995) 449–456. doi:10.3168/jds.S0022-0302(95)76654-1.

[195] C.R. Henderson, Best linear unbiased estimation and prediction under a selection model., Biometrics. 31 (1975) 423–47. http://www.ncbi.nlm.nih.gov/pubmed/1174616 (accessed January 14, 2016).

[196] S.H. Katsanis, N. Katsanis, Molecular genetic testing and the future of clinical genomics, Nat Rev Genet. 14 (2013) 415–426. doi:10.1038/nrg3493.

[197] G. Abraham, M. Inouye, Genomic risk prediction of complex human disease and its clinical application This review comes from a themed issue on Molecular and genetic bases of disease, Curr. Opin. Genet. Dev. 33 (2015) 10–16. doi:10.1016/j.gde.2015.06.005.

[198] S.M. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, P.F. Sullivan, P. Sklar, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder., Nature. 460 (2009) 748–52. doi:10.1038/nature08185.

[199] G. de los Campos, D. Gianola, D.B. Allison, Predicting genetic predisposition in humans: the promise of whole-genome markers., Nat. Rev. Genet. 11 (2010) 880–6. doi:10.1038/nrg2898.

[200] M.E. Goddard, N.R. Wray, K. Verbyla, P.M. Visscher, Estimating Effects and Making Predictions from Genome-Wide Marker Data, Stat. Sci. 24 (2010) 517–529. doi:10.1214/09-STS306.

[201] L.R. Schaeffer, E. Fimland, J.F. Hayes, W.G. Hill, C.R. Henderson, C.R. Henderson, C.R. Henderson, R.L. Quaas, E.J. Pollak, R.L. Quaas, E.J. Pollak, R.L. Quaas, M.F. Rothschild, C.R. Henderson, R.L. Quaas, L.R. Schaeffer, L.R. Schaeffer, L.R. Schaeffer, J.W. Wilton, L.R. Schaeffer, J.W. Wilton, L.R. Schaeffer, J.W. Wilton, R. Thompson, R. Thompson, A.K.W. Tong, B.W. Kennedy, J.E. Moxley, Sire and Cow Evaluation Under Multiple Trait Models, J. Dairy Sci. 67 (1984) 1567–1580. doi:10.3168/jds.S0022-0302(84)81479-4.

[202] R. Thompson, K. Meyer, A review of theoretical aspects in the estimation of breeding

values for multi-trait selection, Livest. Prod. Sci. 15 (1986) 299–313. doi:10.1016/0301-6226(86)90071-0.

[203] B.J. Vilhjálmsson, J. Yang, H.K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E.E. Kenny, M.H. Schierup, P. De Jager, N.A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P.M. Visscher, P. Kraft, N. Patterson, A.L. Price, S. Ripke, B.M. Neale, A. Corvin, J.T.R. Walters, K.-H. Farh, P.A. Holmans, P. Lee, B. Bulik-Sullivan, D.A. Collier, H. Huang, T.H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S.A. Bacanu, M. Begemann, R.A. Belliveau, J. Bene, S.E. Bergen, E. Bevilacqua, T.B. Bigdeli, D.W. Black, R. Bruggeman, N.G. Buccola, R.L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R.M. Cantor, V.J. Carr, N. Carrera, S. V. Catts, K.D. Chambert, R.C.K. Chan, R.Y.L. Chen, E.Y.H. Chen, W. Cheng, E.F.C. Cheung, S.A. Chong, C.R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J.J. Crowley, D. Curtis, M. Davidson, K.L. Davis, F. Degenhardt, J. Del Favero, L.E. DeLisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A.H. Fanous, M.S. Farrell, J. Frank, L. Franke, R. Freedman, N.B. Freimer, M. Friedl, J.I. Friedman, M. Fromer, G. Genovese, L. Georgieva, E.S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J.I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. de Haan, C. Hammer, M.L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A.M. Hartmann, F.A. Henskens, S. Herms, J.N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D.M. Hougaard, M. Ikeda, I. Joa, A. Julia, R.S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M.C. Keller, B.J. Kelly, J.L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J.A. Knowles, B. Konte, V. Kucinskas, Z.A. Kucinskiene, H. Kuzelova-Ptackova, A.K. Kahler, C. Laurent, J.L.C. Keong, S.H. Lee, S.E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C.M. Loughland, J. Lubinski, J. Lnnqvist, M. Macek, P.K.E. Magnusson, B.S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R.W. McCarley, C. McDonald, A.M. McIntosh, S. Meier, C.J. Meijer, B. Melegh, I. Melle, R.I. Mesholam-Gately, A. Metspalu, P.T. Michie, L. Milani, V. Milanova, Y. Mokrab, D.W. Morris, O. Mors, P.B. Mortensen, K.C. Murphy, R.M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D.A. Nertney, G. Nestadt, K.K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan,

C. O'Dushlaine, F.A. O'Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G.N. Papadimitriou, S. Papiol, E. Parkhomenko, M.T. Pato, T. Paunio, M. Pejovic-Milovancevic, D.O. Perkins, O. Pietilinen, J. Pimm, A.J. Pocklington, J. Powell, A. Price, A.E. Pulver, S.M. Purcell, D. Quested, H.B. Rasmussen, A. Reichenberg, M.A. Reimers, A.L. Richards, J.L. Roffman, P. Roussos, D.M. Ruderfer, V. Salomaa, A.R. Sanders, U. Schall, C.R. Schubert, T.G. Schulze, S.G. Schwab, E.M. Scolnick, R.J. Scott, L.J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J.M. Silverman, K. Sim, P. Slominsky, J.W. Smoller, H.-C. So, C.C.A. Spencer, E.A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R.E. Straub, E. Strengman, J. Strohmaier, T.S. Stroup, M. Subramaniam, J. Suvisaari, D.M. Svrakic, J.P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P.A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B.T. Webb, M. Weiser, D.B. Wildenauer, N.M. Williams, S. Williams, S.H. Witt, A.R. Wolen, E.H.M. Wong, B.K. Wormley, J.Q. Wu, H.S. Xi, C.C. Zai, X. Zheng, F. Zimprich, N.R. Wray, K. Stefansson, P.M. Visscher, R. Adolfsson, O.A. Andreassen, D.H.R. Blackwood, E. Bramon, J.D. Buxbaum, A.D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C.M. Hultman, N. Iwata, A. V. Jablensky, E.G. Jonsson, K.S. Kendler, G. Kirov, J. Knight, T. Lencz, D.F. Levinson, Q.S. Li, J. Liu, A.K. Malhotra, S.A. McCarroll, A. McQuillin, J.L. Moran, P.B. Mortensen, B.J. Mowry, M.M. Nthen, R.A. Ophoff, M.J. Owen, A. Palotie, C.N. Pato, T.L. Petryshen, D. Posthuma, M. Rietschel, B.P. Riley, D. Rujescu, P.C. Sham, P. Sklar, D. St. Clair, D.R. Weinberger, J.R. Wendland, T. Werge, M.J. Daly, P.F. Sullivan, M.C. O'Donovan, P. Kraft, D.J. Hunter, M. Adank, H. Ahsan, K. Aittomäki, L. Baglietto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S.J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. dos S. Silva, D. Easton, A.H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M.M. Gaudet, L. Gibson, C.A. Haiman, P. Hall, A. Hazra, R. Hein, B.E. Henderson, A. Hofman, J.L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M.G. Kibriya, P. Lichtner, S. Lindström, J. Liu, E. Lund, E. Makalic, A. Meindl, H. Meijers-Heijboer, B. Müller-Myhsok, T.A. Muranen, H. Nevanlinna, P.H. Peeters, J. Peto, R.L. Prentice, N. Rahman, M.J. Sánchez, D.F. Schmidt, R.K. Schmutzler, M.C. Southey, R. Tamimi, R. Travis, C. Turnbull, A.G. Uitterlinden, R.B. van der Luijt, Q. Waisfisz, Z. Wang, A.S. Whittemore, R. Yang, W. Zheng, Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores, Am. J. Hum. Genet. 97 (2015) 576–592. doi:10.1016/j.ajhg.2015.09.001.

[204] R.M. Cantor, K. Lange, J.S. Sinsheimer, Prioritizing GWAS Results: A Review of

Statistical Methods and Recommendations for Their Application, Am. J. Hum. Genet. 86 (2010) 6–22. doi:10.1016/j.ajhg.2009.11.017.

[205] C.R. Henderson, Sire evaluation and genetic trends, Proc. Anim. Breed. Genet. Symp. Honour J.L. Lush. (1973) 10–41.

[206] C.R. Henderson, R.L. Quaas, Multiple trait evaluation using relatives' records, J. Anim. Sci. (1976). http://agris.fao.org/agris-search/search.do?recordID=US19770150257 (accessed January 14, 2016).

[207] H.F. Smith, A discrimant function for plant selection, Ann. Eugen. 7 (1936) 240–250.

[208] L.N. Hazel, J.L. Lush, The Efficiency of Three Methods of Selection, J. Hered. 33 (1942) 393–399. http://jhered.oxfordjournals.org/content/33/11/393%5Cnhttp://jhered.oxfordjournals.org.ezproxy.library.wisc.edu/content/33/11/393.extract.

[209] L.N. Hazel, The Genetic Basis for Constructing Selection Indexes., Genetics. 28 (1943) 476–490.

[210] Y.C.J. Wientjes, P. Bijma, R.F. Veerkamp, M.P.L. Calus, An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments, Genetics. 202 (2016) 799–823. doi:10.1534/genetics.115.183269.

[211] N.R. Wray, J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, P.M. Visscher, Pitfalls of predicting complex traits from SNPs., Nat. Rev. Genet. 14 (2013) 507–15. doi:10.1038/nrg3457.

[212] S.H. Lee, W.M.S.P. Weerasinghe, N.R. Wray, M.E. Goddard, J.H.J. Van Der Werf, Using information of relatives in genomic prediction to apply effective stratified medicine, Nat. Publ. Gr. (2017) 1–13. http://dx.doi.org/10.1038/srep42091.

[213] M. Goddard, Genomic selection: prediction of accuracy and maximisation of long term response, Genetica. (2006) 245–257.

[214] C.A. Rietveld, T. Esko, G. Davies, T.H. Pers, P. Turley, B. Benyamin, F. Christopher, V. Emilsson, A.D. Johnson, J.J. Lee, C. De Leeuw, R.E. Marioni, S.E. Medland, M.B. Miller, O. Rostapshova, S.J. Van Der Lee, A.A.E. Vinkhuyzen, N. Amin, D. Conley,

C.M. Van Duijn, R. Fehrmann, L. Franke, E.L. Glaeser, N.K. Hansell, C. Hayward, W.G. Iacono, C. Ibrahim-verbaas, V. Jaddoe, D. Laibson, P. Lichtenstein, C. David, P.K.E. Magnusson, N.G. Martin, G. Mcmahon, N.L. Pedersen, S. Pinker, D.J. Porteous, D. Posthuma, F. Rivadeneira, B.H. Smith, J.M. Starr, H. Tiemeier, J. Nicholas, M. Trzaskowski, A.G. Uitterlinden, C. Frank, M.E. Ward, M.J. Wright, G. Davey, I.J. Deary, M. Johannesson, R. Plomin, M. Peter, D.J. Benjamin, D. Cesarini, D. Philipp, Common genetic variants associated with cognitive performance identified using the proxy-phenotype method, 24 (2015) 2938–2939. doi:10.1073/pnas.1424631112.

[215] Y. Banda, M.N. Kvale, T.J. Hoffmann, S.E. Hesselson, D. Ranatunga, H. Tang, C. Sabatti, L.A. Croen, B.P. Dispensa, M. Henderson, C. Iribarren, E. Jorgenson, L.H. Kushi, D. Ludwig, D. Olberg, C.P. Quesenberry, S. Rowell, M. Sadler, L.C. Sakoda, S. Sciortino, L. Shen, D. Smethurst, C.P. Somkin, S.K. Van Den Eeden, L. Walter, R.A. Whitmer, P.-Y. Kwok, C. Schaefer, N. Risch, Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort, Genetics. 200 (2015) 1285–1295. doi:10.1534/genetics.115.178616.

[216] P. Sklar, S. Ripke, L.J. Scott, O.A. Andreassen, S. Cichon, N. Craddock, H.J. Edenberg, Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4., Nat. Genet. 43 (2011) 977–83. doi:10.1038/ng.943.

Thesis appendix: A practical introduction to some theoretical concepts in quantitative genetics

Introduction

The purpose of this document is to give an overview over frequently occurring quantities in quantitative genetics and to demonstrate how to they can be validated in R examples using simulated genotype data.

It is intended as a guide for students in the field of quantitative genetics and focuses on the demonstration of theoretical concepts through simulation and the highlighting of connections between related concepts. Many concepts are therefore presented in a simplified way.

It has been originally created as an html document and is here reproduced without some of the dynamic features of the original document.

Table 18: Notation

Symbol	Meaning
$\mathbf{X}_{012}$	Original genotype matrix
$\mathbf{X}^*$	Mean-centered genotypes
$\mathbf{X}$	Scaled genotypes (each SNP has mean 0, variance 1)
$\mathbf{y}$	Phenotype
p	Minor Allele Frequency (MAF)
M	Number of SNPs
N	Number of individuals
M_e	Effective number of SNPs
N_e	Effective number of individuals
$\mathbf{A}$	Genetic relatedness matrix (GRM)
$\mathbf{V}$	Variance-covariance matrix of the phenotype
$\mathbf{I}_M$	Identity matrix of dimension $M \times M$
β^*	Effect of a SNP (assumes unscaled genotypes)
β	Effect of a SNP (assumes scaled genotypes)
g	Genetic effect of an individual
i	Index for individuals
j	Index for SNPs
σ_y^2	Phenotypic variance. Often assumed to be 1. $\sigma_g^2 + \sigma_e^2$
σ_g^2	Genetic variance
σ_e^2	Error variance
h^2	SNP-heritability. $\frac{\sigma_g^2}{\sigma_y^2}$
$\hat{p}$	Estimate of p
$\overline{p}$	Mean of p
$var(\mathbf{y})$	(Scalar) variance of y (usually 1 here)
$Var[\mathbf{y}]$	Variance-covariance matrix of the random variable $\mathbf{y}$

Table 19: Summary of equations

Quantity	Definition
Genotype properties	
MAF estimate for a SNP j	$\hat{p}_j = \frac{\overline{\mathbf{X}}_{012,j}}{2}$
Expected MAF sampling variance	$SE^2_{\hat{p}_j} = var(\hat{p}_j \mid p_j) = \frac{\hat{p}_j\,(1-\hat{p}_j)}{2N}$
Expected variance of a SNP	$var(\mathbf{X}_{012,j}) = 2\,\hat{p}_j\,(1-\hat{p}_j)$
LD matrix	$\mathbf{L} = \frac{\mathbf{X}^\top\mathbf{X}}{N}$
GRM	$\mathbf{A} = \frac{\mathbf{X}\mathbf{X}^\top}{M}$
LD score of a SNP	$l_j = \frac{1}{N^2}\mathbf{X}_j^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}_j$
SNP effect estimates	
OLS effect estimate for model with one SNP ($\hat{\beta}_{GWAS}$)	$\hat{\beta}_{j,GWAS} = \frac{\mathbf{X}_j^\top\mathbf{y}}{\mathbf{X}_j^\top\mathbf{X}_j} = \frac{cov(\mathbf{X}_j,\mathbf{y})}{var(\mathbf{X}_j)}$
Mixed linear mode association (MLMA) estimate for one SNP	$\hat{\beta}_{j,MLMA} = \frac{\mathbf{X}_j^\top\mathbf{V}^{-1}\mathbf{y}}{\mathbf{X}_j^\top\mathbf{V}^{-1}\mathbf{X}_j}$
OLS effect estimate for model with all SNPs ($\hat{\boldsymbol{\beta}}_{OLS}$)	$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
BLUP effect estimate	$\hat{\boldsymbol{\beta}}_{BLUP} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$
Precision of SNP effect estimates	
Expected sampling variance of $\hat{\beta}^*_{j,GWAS}$	$SE^2_{\hat{\beta}^*_j} = var(\hat{\beta}^*_j \mid \beta^*_j) \approx \frac{1}{N \times var(\mathbf{X}_j)}$
Expected sampling variance of $\hat{\beta}_{j,GWAS}$	$SE^2_{\hat{\beta}_j} = var(\hat{\beta}_j \mid \beta_j) \approx \frac{1}{N}$
Expected variance of $\hat{\beta}_{j,GWAS}$ assuming independent markers	$var(\hat{\beta}_j) = var(\beta_j) + var(\hat{\beta}_j \mid \beta_j)$ $= \frac{h^2}{M} + \frac{1}{N}$
Expected variance of $\hat{\beta}_{j,GWAS}$	$var(\hat{\beta}_j) = \frac{h^2}{M_e} + \frac{1}{N}$

Expected accuracy of GWAS predictor	$cor^2(\mathbf{y}, \hat{\mathbf{g}}_{GWAS}) = \frac{h^2}{1 + \frac{M_e}{Nh^2}}$
Expected accuracy of BLUP predictor	$cor^2(\mathbf{y}, \hat{\mathbf{g}}_{BLUP}) = R^2 = \frac{h^2}{1 + \frac{M_e(1-R^2)}{Nh^2}}$

The structure of genotype data

Simulating genotypes

Humans have diploid genomes, so at each biallelic SNP, there are 2×2 possible combinations of alleles at each locus. Since we don't usually distinguish between, say AG and GA, we are left with 3 distinct genotypes, which means that we can code genotypes for each SNP and individual as 0, 1 or 2, which can be interpreted as 0, 1 or 2 alternative alleles.

Under a random mating assumption, the number of alternative alleles for a SNP and individual follows a binomial distribution with 2 draws (one from mum and one from dad) and probability equal to the minor allele frequency of that SNP.

Let's assume that the minor allele frequency of our M SNPs come from a uniform distribution between 0 and 0.5.

```
set.seed(6155)
m = 500                                  # number of SNPs
maf = runif(m, 0, .5)                    # random MAF for each SNP
# apologies if using "=" as the assignment operator in R makes your eyes
hurt
```

The `set.seed` command here ensures that the random draws from a distribution will be the same each time this code is run. We can then draw genotypes for one person for each SNP.

```
x012 = rbinom(m, 2, maf)
```

We want genotypes for N individuals, so let's replicate this N times.

```
n = 400                                    # number of individuals
x012 = t(replicate(n, rbinom(m, 2, maf)))  # n x m genotype matrix
```

If we are unlucky, some SNPs will be monomorphic, so not actually vary between people. Since this can cause problems, let's simulate more SNPs and only keep M polymorphic ones.

```
x012 = t(replicate(n, rbinom(2*m, 2, c(maf, maf))))
polymorphic = apply(x012, 2, var) > 0
x012 = x012[,polymorphic][,1:m]
maf = c(maf, maf)[polymorphic][1:m]
round(maf[1:10], 2)
```

```
##  [1] 0.02 0.04 0.28 0.18 0.08 0.10 0.45 0.03 0.22 0.26
x012[1:5, 1:10]
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    0    0    0    0    0    1    0    0     0
## [2,]    0    0    1    0    0    0    2    0    0     0
## [3,]    0    0    0    1    0    0    0    0    1     1
## [4,]    0    1    1    1    0    0    1    0    1     1
## [5,]    0    0    1    0    0    0    1    0    1     0
```

Later on, we will not only need the original genotype matrix $\mathbf{X}_{012}$, but also a version in which each SNP is mean-centered, $\mathbf{X}^*$, and a version in which each SNP has mean 0 and variance 1, $\mathbf{X}$.

```
x = scale(x012, scale=FALSE)                    # mean 0
x01 = scale(x012, scale=TRUE)                   # mean 0, variance 1
# note that the x in R is X* in the text and x01 in R is X in the text,
similar with beta.
```

Working with mean-centered genotypes and phenotypes makes life a lot easier, but it means we don't have to estimate intercept terms. So a model like this:

$$\mathbf{y} \sim \beta_0 + \beta_1 \mathbf{x} + \mathbf{e}$$

just becomes

$$\mathbf{y} \sim \beta \mathbf{x} + \mathbf{e}$$

Similarly, assuming that there are no covariates, or that y has already been corrected for any covariates, makes things a lot simpler as well.

MAF estimate

Here, we know what the true MAF is for each SNP (p), because we have simulated it, but in real data we will have to get an estimate based on a finite sample. For a diploid genome coded in 0 and 1, this estimate would just be the mean across individuals, but since we have a diploid genome, it is the mean divided by two.

$$\hat{p}_j = \frac{\overline{\mathbf{x}}_{012,j}}{2}$$

```
maf_est = colMeans(x012)/2
qplot(maf, maf_est, col=maf) +
  geom_abline() + xlab('p') + ylab(expression(hat(p)))
```

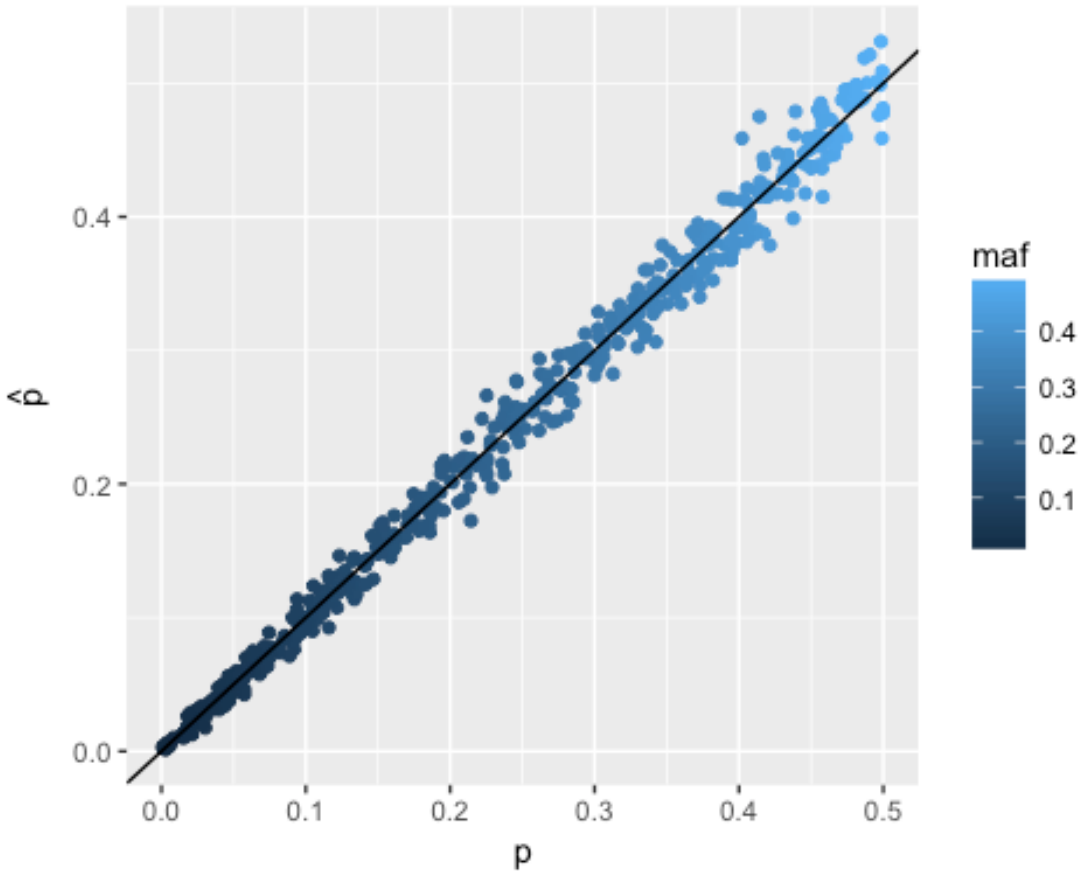

Figure 41: True and estimated minor allele frequency

Strictly speaking, we are estimating the frequency of the alternative allele (the one coded as 2 when homozygous), but since we simulated MAFs in the range [0, 0.5], we can talk about the minor allele frequency.

Sampling variance of MAF estimates

For most SNPs we get a reasonable MAF estimate. If we want to dig deeper, we can also quantify how close our MAF estimate is, on average, to the true value. So we want to know what $p_j - \hat{p}_j$ is, on average. If we have an unbiased estimate, $p_j - \hat{p}_j$ will be zero on average, but it is also interesting what the variance of this quantity is:

$$SE^2_{\hat{p}_j} = var(\hat{p}_j \mid p_j) = var(\hat{p}_j - p_j)$$

This is called the sampling variance, or the (squared) standard error. Very often when the standard error is mentioned, it is about the standard error of a mean estimate, which is

$$SE^2_{\bar{x}} = \frac{var(x)}{N}$$

However, any kind of estimate has a standard error, and it is not always as straight forward to calculate. In this case here, we want to know the sampling variance of the MAF estimate, which is in fact the standard error of a mean estimate, because the MAF estimate is the mean of the genotype values divided by two. So:

$$SE^2_{\hat{p}_j} = var(\hat{p}_j - p_j) = \frac{var(\frac{\mathbf{X}_j}{2})}{N} = \frac{var(\mathbf{X}_j)}{4N}$$

Since it depends on a finite sample, the standard error is itself an estimate and should get a hat.

$$\hat{SE}^2_{\hat{p}_j} = \frac{var(\mathbf{X}_j)}{4N}$$

But we're short on hats, so some quantities will be missing them even though they should have them.

The variance of a genotype can be estimated as $var(\mathbf{X}_j) = 2p_j(1-p_j)$

So the standard error of $\hat{p}_j$ can be also estimated as

$$\hat{SE}^2_{\hat{p}_j} = \frac{\hat{p}_j(1-\hat{p}_j)}{2N}$$

Let's see how this estimated standard error of $\hat{p}_j$ compares to the actual fluctuation of $\hat{p}_j$ around p_j. For this we have to group SNPs by MAF, because the standard error of $\hat{p}_j$ depends on p_j.

```
ss = .02
sequ = seq(0, max(maf_est), ss)
bins = cut(maf_est, sequ)
binMeans = (sequ + ss/2)[-length(sequ)]
dat = data.frame(observed=tapply(colMeans(x012)/2 - maf, bins, var),
                 expected=binMeans*(1-binMeans)/(2*n),
```

```
                binNum=1:length(binMeans))

ggplot(dat, aes(expected, observed)) +
  geom_text(aes(label=binNum, col=binMeans)) +
  geom_abline() +
  xlab(expression(paste(hat(SE)[hat(p)]^2, "  (",hat(var),"(", hat(p), "
| p))"))) +
  ylab(expression(paste("var(",hat(p)," | p)")))
```

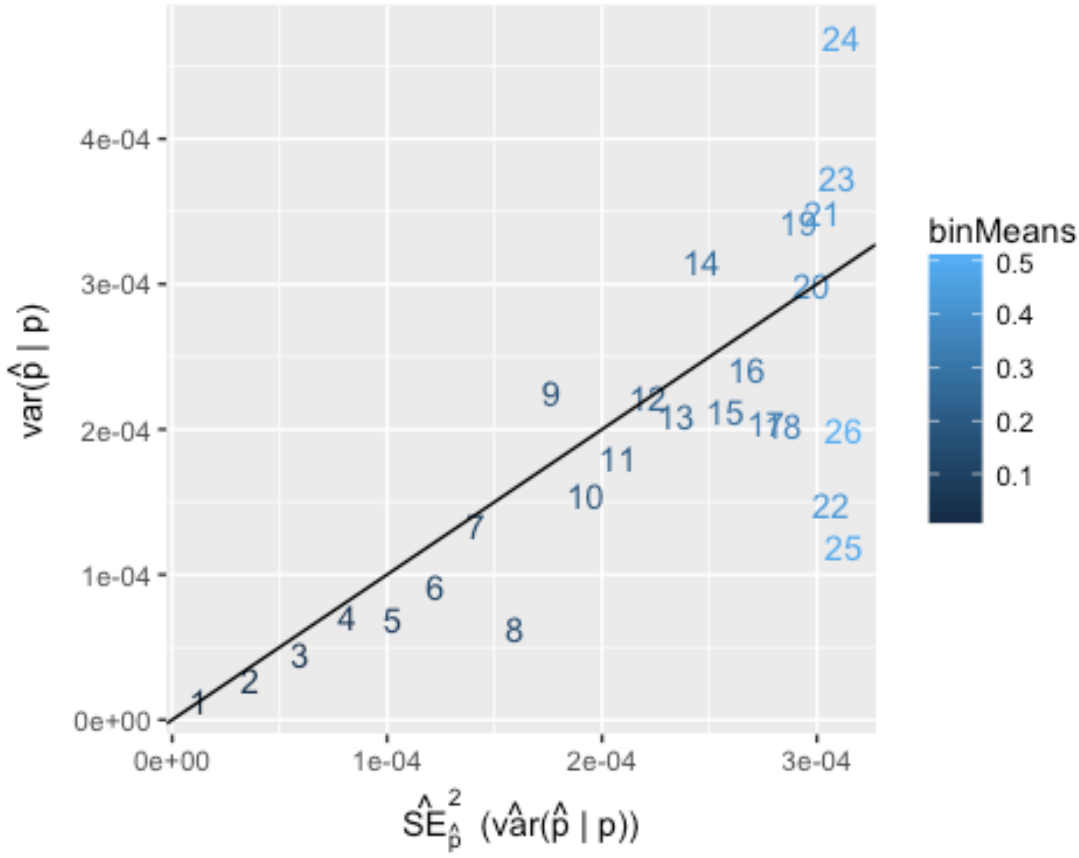

Figure 42: MAF standard error estimate vs actual variance of the MAF estimate

So for the higher MAF bins the empirical $var(\hat{p}_j - p_j)$ is not very well estimated by our estimate of $SE^2_{\hat{p}_j}$. In other words, the standard error of the standard error estimate of $\hat{p}_j$ gets larger with p_j. Determining $\hat{SE}^2_{\hat{SE}^2_{\hat{p}_j}}$, or $var(\hat{var}(\hat{p}_j \mid p_j) \mid var(\hat{p}_j \mid p_j))$ is left as an exercise for the reader.

Variance of a genotype

I mentioned before that the variance of a genotype for a SNP can be estimated as

$$var(x) = 2\,\hat{p}\,(1-\hat{p})$$

This is simply the expected variance of a binomially distributed random variable.

At the same time this is the expected frequency of heterozygous genotypes under Hardy-Weinberg equilibrium.

```
varx = apply(x012, 2, var)
p1 = qplot(2*maf*(1-maf), varx, col=maf) + geom_abline()
# genotype frequencies
p2 = ggplot(data.frame(x=c(0, 1)), aes(x)) +
  stat_function(fun=function(x) 2*x*(1-x), col='red') +
  stat_function(fun=function(x) x^2, col='green') +
  stat_function(fun=function(x) (1-x)^2, col='blue') +
  ylim(c(0,1)) +
  xlab('allele frequency') +
  ylab('genotype frequencies / variance of x (red)')
grid.arrange(p1, p2, ncol=2)
```

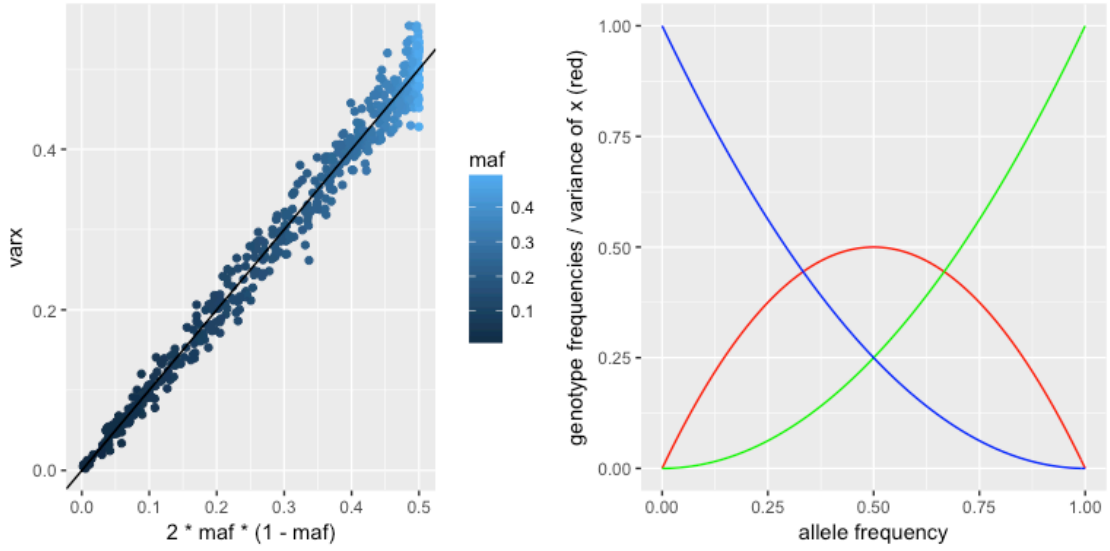

Figure 43: Genotype variance

Left: True genotype variance vs estimate of 2p(1-p). Right: Expected genotype frequencies as a function of MAF. The red curve shows the genotype variance / average heterozygosity as a function of p.

Linkage disequilibrium (LD) Matrix

SNPs are often correlated with one another. Especially when they are nearby, since recombination rarely breaks up any correlation between them. This correlation between a pair of SNPs (two columns in $\mathbf{X}$) is called linkage disequilibrium (LD) and can be estimated by calculating the correlation coefficient between the SNP genotypes.

The LD matrix contains the correlations of all SNP pairs in the genotype matrix and has therefore dimensions $M \times M$. Since correlation implicitly scales by the variance,

$$cor(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

it doesn't matter whether the scaled or unscaled genotype matrix is used. Further, when $\mathbf{X}$ is already scaled, the covariance matrix is the same as the correlation matrix, so the LD matrix can be calculated as:

$$\mathbf{L} = \frac{\mathbf{X}^\top \mathbf{X}}{N}$$

The following should all result in approximately the same LD matrices.

```
ld1 = cor(x012)
ld2 = cor(x01)
ld3 = cov(x01)
ld4 = (t(x01) %*% x01) / n
```

The reason why the last one is slightly different is because `cov` (and `var`) in `R` divide by (n - 1) rather than by n (see Bessel's correction).

LD scores

For some applications, such as LD score regression, it is useful to calculate the LD score of a SNP. This is defined as the sum of squared correlations of a SNP j with all other SNPs:

$$l_j = \sum_{k=1}^{M} cor^2(\mathbf{X}_j, \mathbf{X}_k)$$

Since $\mathbf{X}$ is standardized, we can just calculate the sum of the squared sample correlations like this:

$$\tilde{l}_j = \frac{1}{N^2}\mathbf{X}_j^\top \mathbf{X}\mathbf{X}^\top \mathbf{X}_j$$

However, this is not an unbiased estimate. We can correct for the bias like this:

$$l_j = \frac{\tilde{l}_j\, N - M}{N + 1}$$

```
ldscores_sample = colSums(ld1^2)
ldscores = (ldscores_sample*n - m) / (n + 1)

qplot(ldscores, ldscores_sample, col=ldscores) +
  xlim(c(min(ldscores), max(ldscores))) +
  ylim(min(ldscores), max(ldscores_sample)) +
  geom_abline() +
  scale_colour_continuous(low='black', high='green')
```

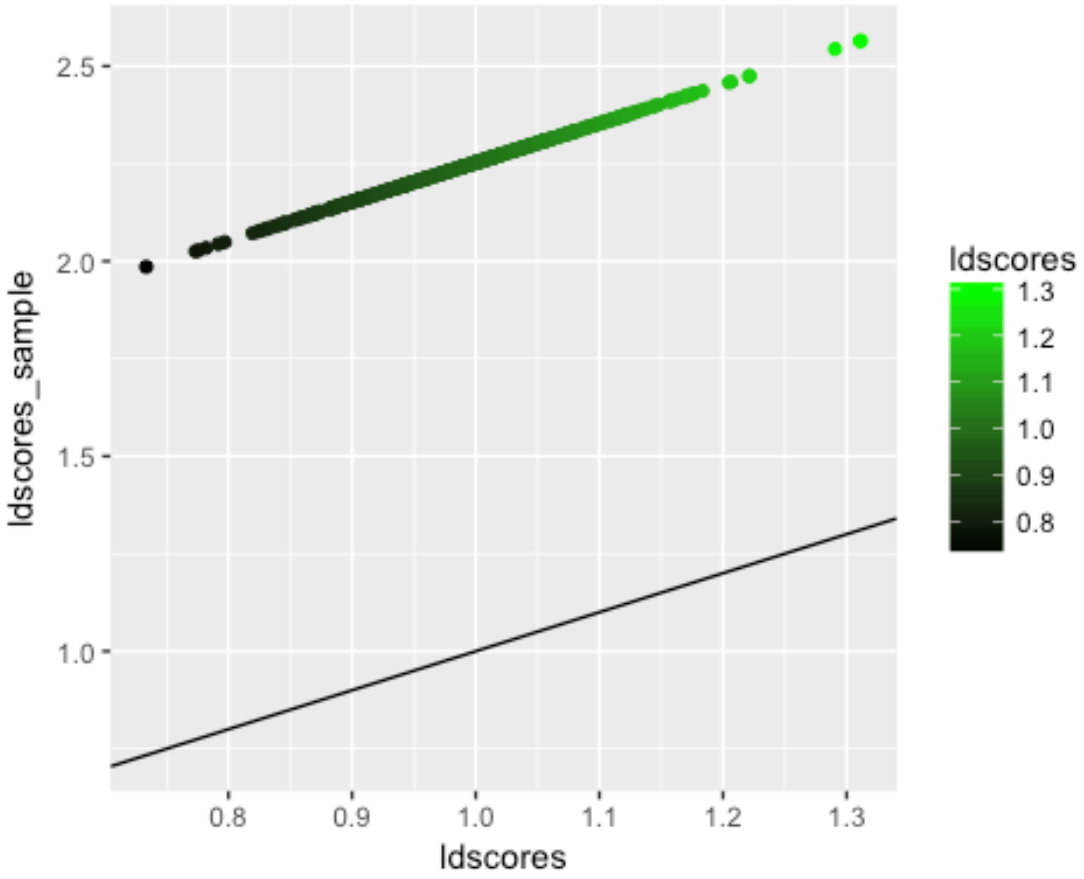

Figure 44: LD scores before and after correcting for biased estimates

In the simulations without LD, LD scores should be centered around one.

Genetic relatedness matrix (GRM)

The genetic relatedness between two individuals (two rows in $\mathbf{X}$) can be estimated as the covariance of the genotypes for these individuals across all SNPs. The GRM contains estimates of the genetic similarity for all pairs of individuals and can be calculated as the covariance matrix of the transposed scaled genotype matrix.

The GRM is actually a genetic similarity matrix. For a genotype in two individuals we want to know if it comes from the same ancestor (identity by descent (IBD)), but we can only measure if it is the same or not (identity by state (IBS)).

It can be calculated as:

$$\mathbf{A} = \frac{\mathbf{X}\mathbf{X}^{\top}}{M}$$

This is very similar to the definition of the LD matrix, but will be a $N \times N$ instead of a $M \times M$ matrix. Also, since $\mathbf{X}$ is usually scaled to have equal variance across SNPs, not across individuals, the diagonal elements on the GRM usually differ from one.

The similarity between the GRM ($\approx \mathbf{X}\mathbf{X}^\top$) and the LD matrix ($\approx \mathbf{X}^\top\mathbf{X}$) explains why models which depend on the GRM often have an equivalent form which depends on the LD matrix.

```
grm = (x01 %*% t(x01))/m
grm2 = cov(t(x01))

p1 = ggplot(melt(grm), aes(value)) + geom_histogram(bins=100)
p2 = ggplot(melt(grm), aes(value)) + geom_histogram(bins=100) +
  coord_cartesian(ylim=c(0,n)) + theme(panel.background=element_blank())
grid.arrange(p1, p2, layout_matrix=matrix(c(1,1,2,1), 2))
```

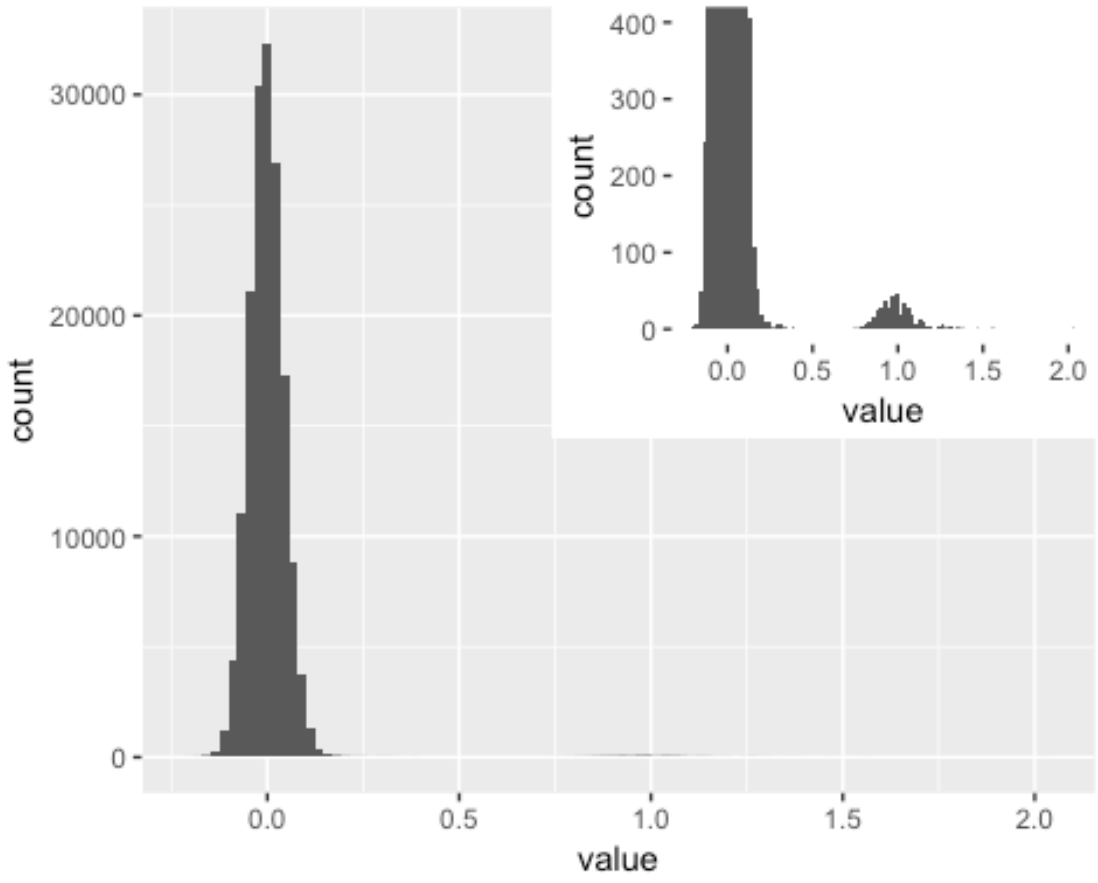

Figure 45: Histogram of GRM values

Inset: Truncated y-axis which highlights the diagonal values centered around one.

Offdiagonal elements represent the relatedness between two individuals (in histogram the large peak centered at 0). Diagonal elements represent the relatedness of an individual with itself, which is the average homozygosity or the level of inbreeding (in histogram the small peak centered at 1). While the LD of a SNP with itself is 1 by definition, the genetic relatedness of an individual with itself can vary around one, because the genotype scaling is performed per SNP, not per individual.

Genetic principal components

The GRM can be used to calculate genetic principal components via singular value decomposition (SVD) of the mean-centered genotype matrix or via eigendecomposition of the GRM. (Comparison of Eigendecomposition and SVD)

Principal component analysis (PCA) can be used to rotate a matrix in such a way, that each column (principal component) is orthogonal to each other column and the columns are sorted by the amount of variance explained. In genome wide genetic data, the first principal components usually capture ancestry or population stratification and can therefore be used to correct for these confounding factors.

```
xt = scale(t(x01), scale=F) # transpose and mean center across
individuals

principal_axes_eigen        = eigen(grm2)$vectors
principal_axes_svd          = svd(xt)$v
principal_axes_prcomp       = prcomp(xt)$rotation

principal_components_eigen  = xt %*% eigen(grm2)$vectors
principal_components_svd1   = xt %*% svd(xt)$v
principal_components_svd2   = svd(xt)$u %*% diag(svd(xt)$d)
principal_components_prcomp = prcomp(xt)$x
```

The above shows different ways to calculate genetic principal components. While the results are the same, the runtime can differ greatly.

```
cor(principal_components_eigen[,1], principal_components_eigen[,2])
## [1] 9.158365e-18
```

All principal components are orthogonal to each other, so uncorrelated.

```
p1 = qplot(1:min(n,m), prcomp(xt)$sdev^2) + xlab('rank') +
ylab('eigenvalues')
p2 = qplot(principal_components_eigen[,1],
principal_components_eigen[,2]) +
  xlab('PC1') + ylab('PC2')
grid.arrange(p1, p2, ncol=2)
```

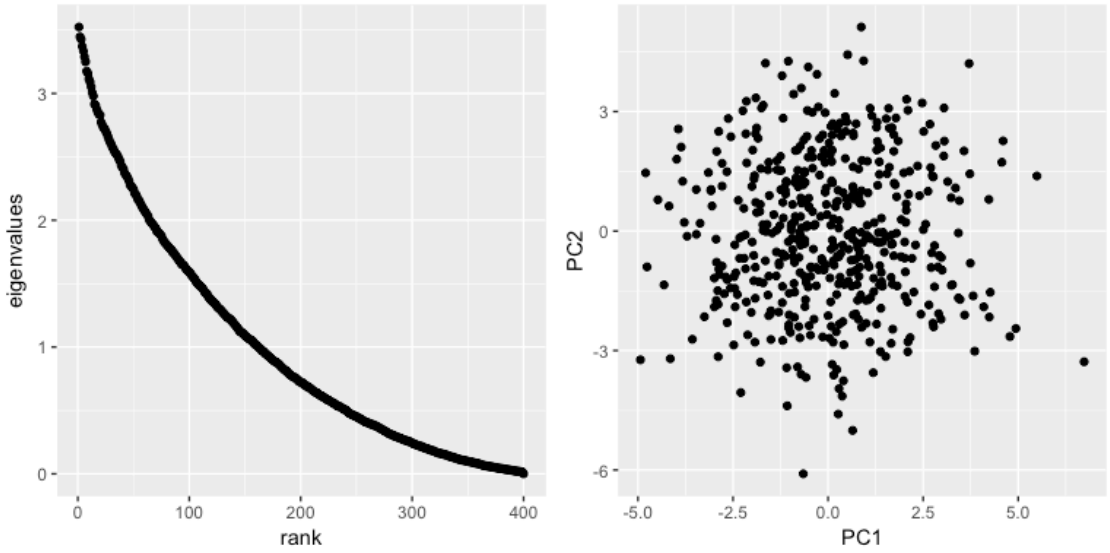

Figure 46: Genotype principal components

Left: The eigenvalues of the GRM indicate what proportion of variance in the genotype is explained by the corresponding genetic principal component. Eigenvectors are sorted by the order of their eigenvalues, so the first principal components always explain more variance than the subsequent ones. Right: No structure in the simulated genotype data.

Simulating genotypes with LD

Real genotype data has LD, and to demonstrate some concepts, we have to briefly leave our cozy fantasy world of only independent SNPs.

```
x012ld = jitter(x012[,rep(1:m, 1:m)[1:m]], .03)
x01ld = scale(x012ld)
ldld = (t(x01ld) %*% x01ld)/n
grmld = (x01ld %*% t(x01ld))/m
ldscores_sampleld = colSums(ldld^2)
ldscoresld = (ldscores_sampleld*n - m) / (n + 1)
```

```
varxld = apply(x012ld, 2, var)


greens = colorRampPalette(c('black', 'green'))(12)
par(mfrow=c(1,2))
image(ld1, col=greens)
image(ldld, col=greens)
```

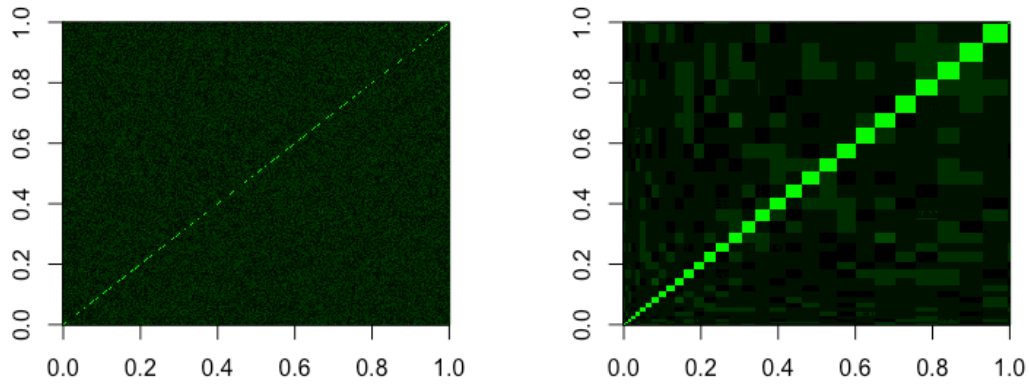

Figure 47: LD matrix in data without and with LD

```
par(mfrow=c(1,1))
```

Effective number of SNPs

When there is LD between SNPs, it is useful to have a quantity that describes how many independent SNPs there are. This is the effective number of SNPs, markers or chromosome segments (M_e), and can be defined as

$$M_e = \frac{M}{\bar{\mathbf{l}}}$$

where $\bar{\mathbf{l}}$ denotes the mean LD score across all SNPs.

In practice, LD scores are often calculated based on a limited number of SNPs (for example 2000 kb or 1 centimorgan), which makes them smaller, so they can't be used to calculate M_e.

Because of the similarity of $\mathbf{A}$ and $\mathbf{L}$, M_e can also be approximated as

$$M_e \approx \frac{1}{\mathbb{E}[\mathbf{A}_{i,j}^2]} \approx \frac{1}{var(\mathbf{A}_{i,j})}$$

where $\mathbf{A}_{i,j}$ denotes all offdiagonal elements of $\mathbf{A}$.

```
m
## [1] 500
(me = m/mean(ldscores))
## [1] 502.6183
1/var(grm[upper.tri(grm)])
## [1] 514.2901
(meld = m/mean(ldscoresld))
## [1] 24.11521
1/var(grmld[upper.tri(grmld)])
## [1] 24.47277
```

M_e is smaller than the number of LD blocks in the data with LD. This is because larger blocks have higher weight.

In an average European population of unrelated individuals M_e is between 60,000 and 70,000.

Effective population size

The effective population size (N_e) is the number of individuals that an idealized population would have to have to result in the parameters that are being observed in the real population and can be estimated in different ways, depending on what parameters are of interest.

Due to the symmetry of $\mathbf{A}$ and $\mathbf{L}$, N and M, one could expect that it can be simply estimated as $\frac{1}{\mathbb{E}[\mathbf{L}_{i,j}^2]}$, but it is in fact better approximated by:

$$N_e \approx \frac{1}{4c \times \mathbb{E}[\mathbf{L}_{i,j}^2]} \approx \frac{1}{4c \times var(\mathbf{L}_{i,j})}$$

where 4 accounts for the fact that humans have diploid, not haploid genomes and c is the recombination rate. The absence of a term equivalent to c in the estimation of M_e reflects the assumption of random mating and thus an unstructured population.

```
n
## [1] 400
1/var(ld1[upper.tri(ld1)])
## [1] 400.0754
# estimate of recombination rate in this particular case
recomb = sqrt(2*m)/m
1/(var(ldld[upper.tri(ldld)]) * recomb)
## [1] 391.0233
```

Estimates of N_e in human populations range from around 2,500 in European and Asian populations to around 6,000 in African populations, while c is around 0.01 per Mb.

SNP effects - basics

While the previous parts have been fun, quantitative genetics really on becomes interesting when it is studied in combination with one or more phenotypes, for example to estimate the effect that a SNP has on a phenotype. Here I will first describe a model to simulate phenotypes from genotypes, and then talk about different ways to estimate SNP effects.

Modeling a phenotype

A particular phenotype y of an individual i can be modeled as a combination of a genetic component, g, and an error component e:

$$y_i = g_i + e_i$$

Some researchers choose to focus on the e component and distinguish between environmental effects and various sources of error, but we will focus on the g component and model e as a normally distributed random variable which is independent of g.

Here, we set up a model which says that the similarity (or differences) between our N individuals can be partitioned into a genetic component and an environmental / error component.

$$Var[\mathbf{y}] = Var[\mathbf{g} + \mathbf{e}] = Var[\mathbf{g}] + Var[\mathbf{e}]$$

g and **e** are independent by defintion, so the variance of the sum is the sum of the variances.

$Var[\mathbf{y}]$, $Var[\mathbf{g}]$ and $Var[\mathbf{e}]$ are all $N \times N$ variance-covariance matrices of the phenotypes, the genetic effects and the environmental effects.

In contrast, the scalar variances σ_y^2, σ_g^2 and σ_e^2, represent the variances of the phenotype, the total genetic effect and the environmental effect, respectively.

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2$$

So what is Var[e]? If we assume that the error term is independent and identically distributed (i.i.d.) for each individual, then

$$Var[\mathbf{e}] = \sigma_e^2 \mathbf{I}_N$$

where I_N is a $N \times N$ identity matrix.

What about $Var[\mathbf{g}]$? g_j is the genetic part of the phenotype of an individual (also called genetic value or breeding value). The genotype of an individual is comprised of M markers, each of which has a contribution proportional to the general effect size of this marker, β_j, and the number of effect alleles (0, 1 or 2):

$$g_i = \sum^{j} X_{ij}\,\beta_j$$

Written in matrix notation for all individuals, this becomes

$$g = \mathbf{X}^\top \boldsymbol{\beta}$$

So $Var[\mathbf{g}] = Var[\mathbf{X}\boldsymbol{\beta}]$. Part of our model assumption is that we see SNP effects as random (they are drawn from a probability distribution), but genotypes as fixed (even though we also drew genotypes from a probability distribution to simulate them). This allows to write

$$Var[\mathbf{X}\boldsymbol{\beta}] = \mathbf{X} Var[\boldsymbol{\beta}] \mathbf{X}^\top$$

This is just a more general form of the rule $var(ax) = a^2 var(x)$, for a constant a and random variable x. If we assume that the true SNP effects are all i.i.d., and that the variances of all SNP effects add up to σ_g^2, then

$$Var[\boldsymbol{\beta}] = \frac{\sigma_g^2}{M} \mathbf{I}_M$$

and so

$$Var[\mathbf{X}\boldsymbol{\beta}] = \mathbf{X}(\frac{\sigma_g^2}{M} \mathbf{I}_M)\mathbf{X}^\top = \frac{\sigma_g^2}{M} \mathbf{X}\mathbf{X}^\top$$

$\frac{\mathbf{X}\mathbf{X}^\top}{M}$ is the GRM, so

$$Var[\mathbf{g}] = \sigma_g^2 \mathbf{A}$$

Putting it all together, we get

$$Var[\mathbf{y}] = \sigma_g^2 \mathbf{A} + \sigma_e^2 \mathbf{I}_N$$

This is the model on which our phenotype simulations will be based and it is also the model underlying many other quantitative genetics methods.

Limitations of the model

All models are wrong; some models are useful. Most of this document describes how the model specified above is useful, but it is important to keep in mind the extent to which it is wrong.

No non-genetic effects

Apart from the inclusion of fixed effects, which is omitted here for simplicity, the model doesn't account for environmental effects. Specifically, the equation $Var[\mathbf{y}] = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}_N$ states that the phenotypic similarity among individuals is partly due to genetic factors, and that the GRM tells us who should be similar to whom, and partly due to other factors, which are unique for each individual, and independent of their genetic relatedness. This assumption would not be justified when working with close relatives - close relatives share environments as well as genes. It is less problematic in unrelated individuals.

No dominance effects

Another assumption of our model is that for each SNP the effect of having two copies of the alternative allele is twice that of having on copy. In other words, we assume complete additivity and no dominance effects. This is not as bad as it sounds because the dominance effect is only whatever is left after removing the additive effect, and empirical data suggests that this is usually not a lot.

No epistasis

Dominance can be seen as interaction effects of a SNP with itself, and epistasis describes interaction effects among different SNPs. For example, if there is a SNP which leads to a gene knockout, the effect of another SNP which can lead to a knockout of the same gene will depend on the first SNP. It seems at first that these effects should be ubiquitous, but as with dominance effects, empirical evidence does not suggest that all epistatic effect together

explain as much variance as the additive effects alone, partially because the interaction effects are again only what is left after accounting for additive effects.

That's very useful, because the number of interactions among SNPs can grow very rapidly, which makes it hard to test them exhaustively. With M SNPs, there are $\frac{M(M-1)}{2}$ pairwise interactions, and pairwise interactions are only the simplest kind.

Effect size distribution

Although so far we have not specified any distribution of effect sizes, in the next section we will do so by drawing effects from a normal distribution. This makes the model more tractable, but is not a good approximation for traits with a few loci of large effects. However, for many polygenic traits this does not impose a great limitation, even if the true effect size distribution is not exactly normal. There are many tools and methods which try to better model a wide range of traits by assuming a different distribution of SNP effect sizes.

Simulating a phenotype

According to the model described above we will simulate our phenotypes. Here we will assume that all SNPs have an effect and that their effect size follows a normal distribution.

```
h2 = 0.5
beta01 = rnorm(m, 0, sqrt(h2/m))
beta = beta01/sqrt(varx)
g = x %*% beta
# equivalent to g = x01 %*% beta01
e = rnorm(n, 0, sqrt(1-h2))
y = g + e

var(y) # should be around 1
##          [,1]
## [1,] 1.04066
```

Same for the data with LD:

```
gld = x01ld %*% beta01
eld = rnorm(n, 0, sqrt(1-h2))
yld = gld + eld
```

Here we assume that the (SNP-)heritability is 0.5, and that the effect that each SNP has on the (scaled) phenotype assuming scaled genotypes (effect of $\boldsymbol{\beta}$ on $\mathbf{y}$ assuming $\mathbf{X}$) is drawn from a normal distribution with mean 0 and variance $\frac{h^2}{M}$.

The effect of $\boldsymbol{\beta}$ given $\mathbf{X}$ is identical to the effect of $\boldsymbol{\beta}^*$ given $\mathbf{X}^*$:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^*\boldsymbol{\beta}^*$$

because

$$\beta_j = \beta_j^* \times \sqrt{var(\mathbf{X}_j)}$$

and

$$X_{ij} = \frac{X_{ij}^*}{\sqrt{var(\mathbf{X}_j)}}$$

but by drawing $\boldsymbol{\beta}$ from a normal distribution, not $\boldsymbol{\beta}^*$, we ensure that $\boldsymbol{\beta}$ is independent of p, but $\boldsymbol{\beta}^*$ is not. Absolute values of $\boldsymbol{\beta}^*$ will be larger for rare SNPs (low p). This models the expectation from natural selection which predicts that common variants will on average have smaller effects than rare variants. The seemingly innocuous scaling of genotypes therefore has a profound impact on the way in which the phenotype is modeled. There is some debate over whether this model uses the right relation between minor allele frequency and effect size, but the fact remains that it is very convenient to assume that the variance explained per SNP is independent of minor allele frequency.

```
ma = function(x, n=10) stats::filter(x, rep(1/n, n), sides=2)
p1 = qplot(maf, abs(beta01), col=maf) +
  geom_line(data=data.frame(maf=sort(maf),
beta01=ma(abs(beta01)[order(maf)])),
            col='red', size=1) +
  xlab('p') + ylab(expression(beta)) +
  theme(legend.position='none')
```

```
p2 = qplot(maf, abs(beta), col=maf) +
  geom_line(data=data.frame(maf=sort(maf),
beta=ma(abs(beta)[order(maf)])),
            col='red', size=1) +
  xlab('p') + ylab(expression(paste(beta^"*"))) +
theme(legend.position='none')
grid.arrange(p1, p2, ncol=2)
```

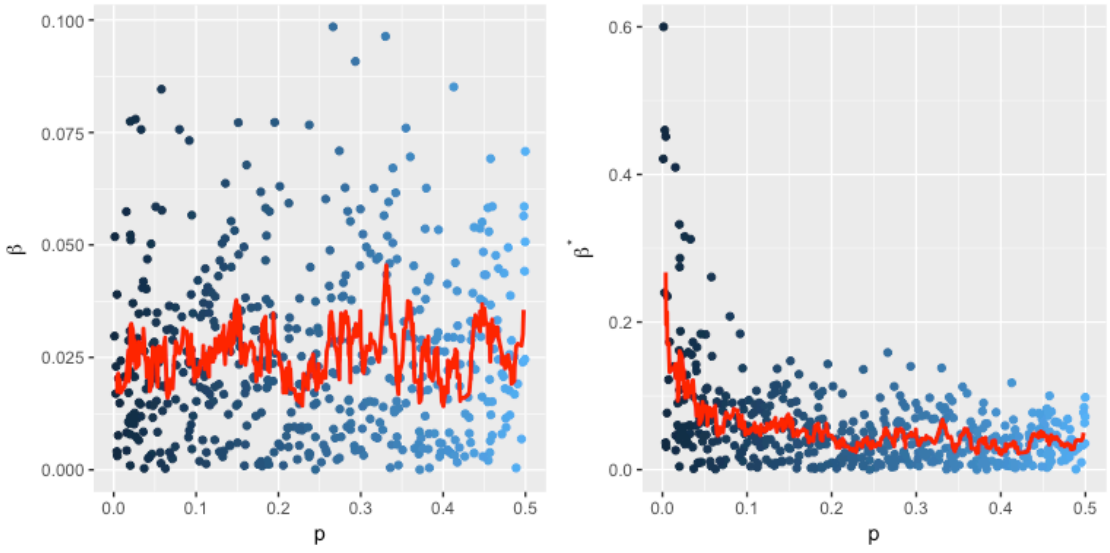

Figure 48: Relation between MAF and effect size

β is independent of MAF, but β^* increases with smaller MAF.

OLS effect estimate for one SNP (simple linear regression, GWAS)

The simplest way to estimate the effect of a SNP on a quantitative trait is simple regression through OLS. Simple regression usually means that it is univariate, or one SNP at a time. This is what is done in GWAS. To contrast this with multivariate regression through OLS, I call the (marginal) estimates from univariate OLS $\hat{\boldsymbol{\beta}}_{GWAS}$ and the (conditional or joint) estimates from multivariate OLS $\hat{\boldsymbol{\beta}}_{OLS}$.

The effect estimate in a simple linear regression is defined as:

$$\hat{\boldsymbol{\beta}}^{*}_{j,GWAS} = \frac{\mathbf{X}_j^{*\top}\mathbf{y}}{\mathbf{X}_j^{*\top}\mathbf{X}_j^{*}}$$

Or for standardized effects:

$$\hat{\boldsymbol{\beta}}_{j,GWAS} = \frac{\mathbf{X}_j^{\top}\mathbf{y}}{\mathbf{X}_j^{\top}\mathbf{X}_j} \approx \frac{\mathbf{X}_j^{\top}\mathbf{y}}{N}$$

```
beta01_gwas = t(x01) %*% y / diag(t(x01)%*%x01)
beta_gwas = t(x) %*% y / diag(t(x)%*%x)

beta01_gwasld = (t(x01ld) %*% yld) / n
beta_gwasld = beta01_gwasld / sqrt(varxld)

p1 = qplot(beta, beta_gwas, col=maf) + geom_abline() +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta, beta_gwas), 3))) +
  theme(legend.position='none')

# compare this with lm
beta_gwas_lm = sapply(1:m, function(i) lm(y ~ x[,i])$coefficients[2])
p2 = qplot(beta_gwas, beta_gwas_lm, col=maf) + geom_abline() +
  theme(legend.position='none')
```

```
grid.arrange(p1, p2, ncol=2)
```

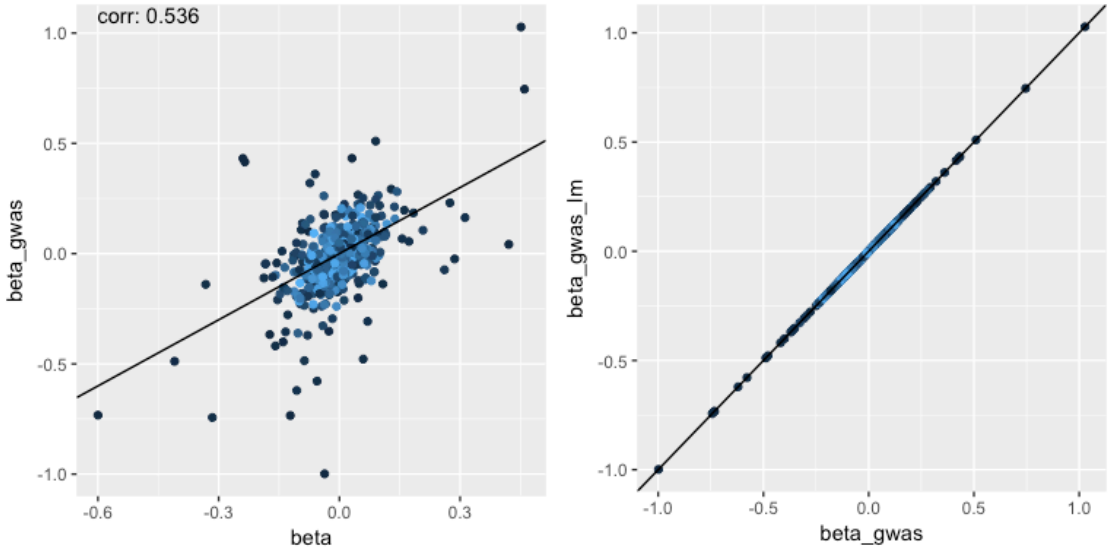

Figure 49: GWAS effect estimates

Left: true SNP effects vs GWAS estimates. Right: two identical ways to estimate the effect in R.

OLS effect estimate for all SNPs (multiple regression, OLS)

Rather than fitting one SNP at a time, we can also fit all SNPs at the same time. This will create effect estimates which are conditional on other SNPs.

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

```
if(n >= m) {
  beta_ols = solve(t(x) %*% x) %*% t(x) %*% y
} else {
    # pseudo beta_ols
  beta_ols =  solve(t(x) %*% x + diag(m)*1e-6) %*% t(x) %*% y
}
p1 = qplot(beta, beta_ols, col=maf) + geom_abline() +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta, beta_ols), 3))) +
  theme(legend.position='none')
```

```
# compare this with lm
p2 = ggplot() + theme(panel.background = element_blank())
if(n >= m) {
  beta_ols_lm = lm(y ~ x)$coefficients[-1]
  p2 = qplot(beta_ols, beta_ols_lm) + geom_abline()
}
grid.arrange(p1, p2, ncol=2)
```

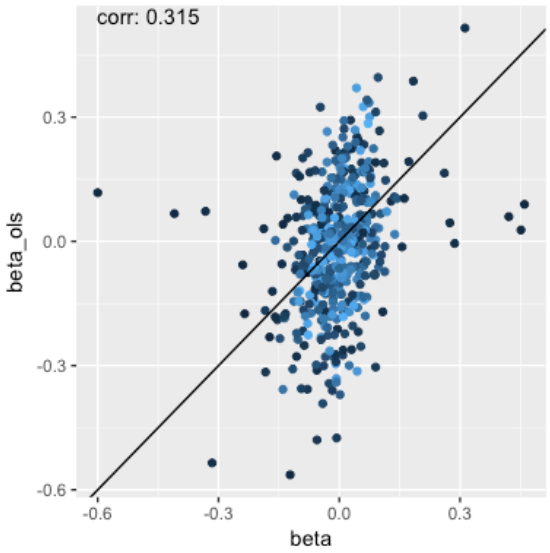

Figure 50: True effects vs multiple regression (OLS) estimates

So what's the difference between the marginal single SNP GWAS model and the conditional multiple SNP OLS model? Imagine two SNPs with large, opposite effect size are in high LD. To the first model, this LD information is inaccessible, and because it can't control for the effect of the other SNP, it will calculate for each SNP an estimate of the combined effect of all the SNPs on the same haplotype block. In this case it means that the effects of both SNP cancel each other out and the GWAS estimate will be very small.

The conditional OLS estimate on the other hand will be able to differentiate between the effects of both SNP, given N is large enough (unless the LD correlation is one). This property of multiple regression can lead to more accurate estimates of the true effects. In our genotype model, all SNPs are independent, so there is not a large difference between conditional and marginal effects. However, the small chance correlation between SNPs that arises through the sampling process is enough to make conditional effect estimates different from marginal effect estimates. On the other hand, conditional effect estimates may not

always be what is desired. Let's assume that a haplotype block with many highly correlated SNPs contains only one SNP with a true, large effect. In a the marginal GWAS model, all correlated SNPs will pick up the signal from this one SNP. In the conditional OLS model, the effect may get smeared out over all correlated SNPs, if N is not large enough to assign the effect to only the causal SNP. By smearing out the effect, none of the SNPs may be found to be significantly associated.

There is also a computational difference between the two models: The multiple regression model is slower because the inversion of $\mathbf{X}^\top\mathbf{X}$ can take time.

Another problem with this model is that it is prone to overfitting the data, as $\frac{M}{N}$ increases. This will lead to less accurate estimates of the true effect size. When $M > N$, it becomes impossible to apply this model, because $\mathbf{X}^\top\mathbf{X}$ will become singular. This problem is being addressed by the BLUP model.

Best Linear Unbiased Prediction (BLUP)

As mentioned before, the multiple regression OLS model above doesn't work if $M > N$. If $M > N$ the system of equations is underdetermined (there are infinitely many solutions for $\hat{\boldsymbol{\beta}}$, $\mathbf{X}^\top\mathbf{X}$ is not positive semidefinite and cannot be inverted). Adding even very small values to the diagonal of $\mathbf{X}^\top\mathbf{X}$ will change that and lead to a unique solution for $\hat{\boldsymbol{\beta}}$. This is called Ridge regression. The shrinkage parameter λ will make estimated effect sizes smaller and should be proportional to the amount of error in the model. In our case, the error is given by the non-genetic component of the phenotype (σ_e or $1 - h^2$ since we assume a phenotypic variance of 1) and λ is defined as

$$\lambda = M\frac{1-h^2}{h^2}$$

While OLS effects for standardized $\mathbf{X}$ and $\hat{\boldsymbol{\beta}}$ are defined as

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

BLUP can be written as

$$\hat{\boldsymbol{\beta}}_{BLUP} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_M)^{-1}\mathbf{X}^\top\mathbf{y}$$

The equation above is equivalent to the following equation which is a function of the GRM, rather than the LD matrix:

$$\hat{\boldsymbol{\beta}}_{BLUP} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{y}$$

This illustrates why it is equivalent to say that BLUP corrects for the LD between SNPs, and that it corrects for the genetic relatedness among individuals.

The last equation can be rearranged to define BLUP as a function of the phenotypic variance-covariance matrix, $\mathbf{V}$:

$$\hat{\boldsymbol{\beta}}_{BLUP} = \frac{h^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$$

where $\mathbf{V} = \sigma_g\mathbf{A} + \sigma_e\mathbf{I}_N = \frac{h^2}{M}(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)$.

Here BLUP is defined for standardized genotypes and SNP effects, because the phenotypic covariance matrix, $\mathbf{V}$, is defined based on standardized genotypes, but conversion between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ is always simple:

$$\hat{\beta}^*_{j,BLUP} \approx \frac{\hat{\beta}_{j,BLUP}}{\sqrt{2p_j(1-p_j)}}$$

```
lambda = m*(1-h2)/h2

beta01_blup = solve(t(x01) %*% x01 + diag(m)*lambda ) %*% t(x01) %*% y
beta_blup = beta01_blup / sqrt(varx)

beta01_blupld = solve(t(x01ld) %*% x01ld + diag(m)*lambda ) %*% t(x01ld)
%*% yld
beta_blupld = beta01_blupld / sqrt(varxld)

p1 = qplot(beta, beta_blup, col=maf) + geom_abline() +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta, beta_blup), 3)))
```

```
p2 = qplot(beta_gwasld, beta_blupld, col=ldscoresld) + geom_abline() +
  scale_colour_continuous(low='black', high='green')
grid.arrange(p1, p2, ncol=2)
```

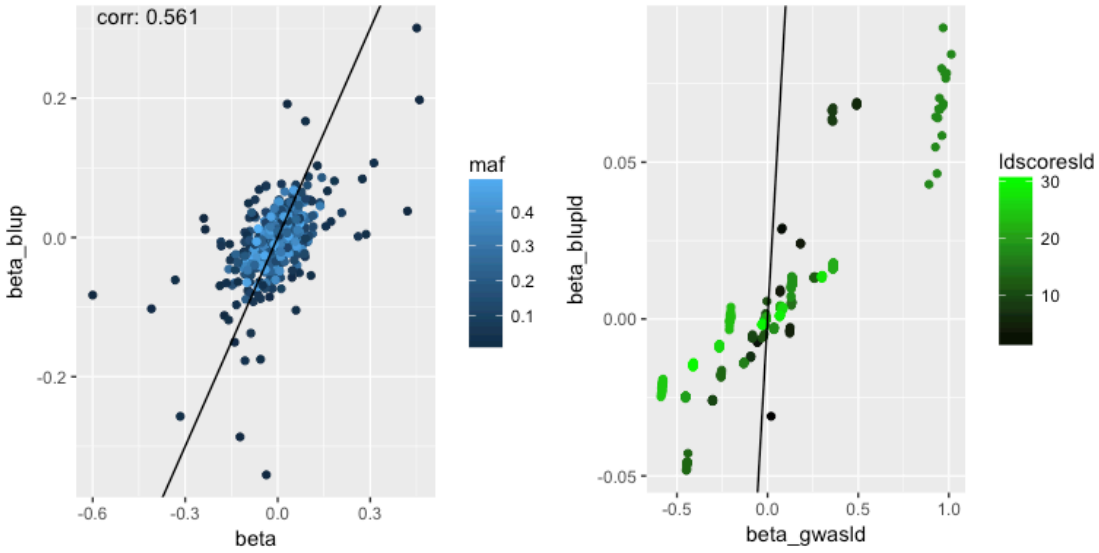

Figure 51: BLUP effect estimates

Left: Data without LD. BLUP estimates usually provide better estimates of the true effect size than GWAS estimates, so the correlation between β and $\hat{\beta}_{BLUP}$ should be a bit larger than between β and $\hat{\beta}_{GWAS}$. Right: Data with LD. Compared to GWAS estimates, BLUP estimates are shrunk, and the shrinkage is proportional to the LD score.

The shrinkage factor λ can vary between 0 and ∞, as h^2 varies between 0 and 1. Consider what happens when h^2 becomes large: λ will go to 0 and $\hat{\boldsymbol{\beta}}_{BLUP}$ will become equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$. The same happens when M decreases (since $\lambda = M\frac{1-h^2}{h^2}$) and also as N increases, since λ will be small relative to the values in the $\mathbf{X}^\top\mathbf{X}$ matrix to which it is added.

What happens in the opposite case, as h^2 and N get smaller and M becomes larger? The matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ will become very heavy along its diagonal, so similar in structure to an identity matrix, but with large values. Consequently, $\hat{\boldsymbol{\beta}}_{BLUP}$ will be shrunk very heavily towards 0, but become very similar to $\hat{\boldsymbol{\beta}}_{GWAS}$.

The importance of the shrinkage factor is not so much that it makes the effects smaller (although this can be useful as well, see winner's curse section). If we just wanted smaller

effects, we could just divide GWAS estimates by some factor, but this wouldn't affect the correlation with the true effects.

The reason why the shrinkage factor is important is because it prevents the estimation procedure from placing too much weight on the correlation structure among SNPs and instead to focus more on the marginal effect of each SNP. This is especially important when M is large relative to N, because this makes it more difficult to estimate purely conditional effects accurately, while the estimability of each marginal effect only depends on N, not M.

Notice that this model assumes that we know what σ_g^2 and σ_e^2 are (or just h^2 with $var(\mathbf{y}) = 1$). In practice, these parameters also have to be estimated from the data at the same time (see variance component estimation section).

A practical tip for calculating BLUP solutions in R: Matrix multiplications are not commutative, but they are associative. That means the order can't change, but we can group them however we like, without changing the result. We can use this to greatly speed up the calculation of chained matrix multiplications, by forcing R to first evaluate the matrix - vector multiplications:

```
invmat = solve(t(x01) %*% x01 + diag(m)*lambda )

system.time( invmat %*% t(x01) %*% y )
##    user  system elapsed
##   0.083   0.002   0.089
system.time( invmat %*% (t(x01) %*% y) )
##    user  system elapsed
##   0.002   0.000   0.001
```

BLUP only has an advantage over OLS in two or more dimensions:

```
if(is.null(out_type)) out_type = ''
if(out_type != 'html') knit_hooks$set(rgl = hook_rgl, webgl = hook_webgl)
sam = 5
x1 = rbinom(sam, 2, .45)
x2 = rbinom(sam, 2, .43)
z = rnorm(sam) - x2
dat = data.frame(x1, x2 , z)
```

```
mfrow3d(nr = 1, nc = 2, sharedMouse = TRUE)
scatter3d(z ~ x1 + x2, data=dat, xlab='x1', ylab='y', zlab='x2',
axis.scales=F,
          fill=out_type=='html', fit='linear')
scatter3d(z ~ x1 + x2, data=dat, xlab='x1', ylab='y', zlab='x2',
axis.scales=F,
          fill=out_type=='html', fit='ridge', lambda=5)
fn = spin3d(axis = c(0, 1, 0))
rglwidget() %>%
  playwidget(controls=par3dinterpControl(fn, 0, 100, steps = 100,
            subscene=subsceneList()[[1]]), step=0, components=c('Play'))
%>%
  playwidget(controls=par3dinterpControl(fn, 0, 100, steps = 100,
            subscene=subsceneList()[[2]]), step=0, components=c('Play'))
```

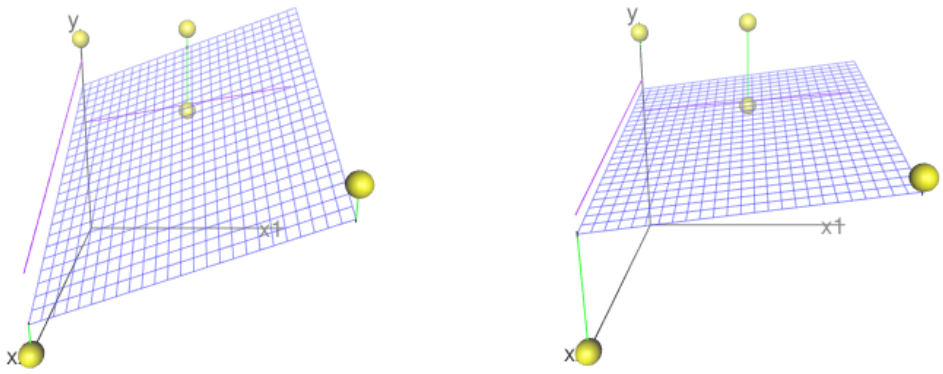

Figure 52: Left: OLS; right: BLUP

Each dot represents an individual with two SNPs (x1 and x2) and a phenotype (y). The surface on the left is the OLS fit to the data, which minimizes the sum of the squared residuals. The slope along the dimensions x1 and x2 represents $\hat{\beta}_1$ and $\hat{\beta}_2$. Here, N is large relative to M (2). If that were not the case, the OLS model could easily overfit the data. To prevent that, BLUP softens the objective of minimizing the squared residuals (green segments), by modeling the y-values not as fixed values, but as measurements with a normally distributed error. The BLUP objective function is then to minimize the sum of

squared residuals, while at the same time limiting the estimated beta values (preventing the surface from becoming too steep). The purple lines are the marginal effects for the two SNPs. On the left, they are just GWAS effects and on the right, they are shrunk marginal effects. BLUP effects should be closer to the marginal effects than OLS effects are to marginal effects.

Mixed linear model association (MLMA)

In a standard GWAS setting, the first 10 or 20 principal components are often fitted as covariates to correct for population stratification. A different approach is to run a mixed linear model association analysis, which is sometimes said to be similar to fitting all principal components, as it fits the whole GRM, which contains the same information as all principal components.

While GWAS estimates are based on ordinary least squares (OLS), MLMA estimates are based on generalized least squares (GLS).

The GWAS association statistic is based on univariate OLS and is

$$\hat{\beta}_{j,GWAS} = \frac{\mathbf{X}_j^\top \mathbf{y}}{\mathbf{X}_j^\top \mathbf{X}_j}$$

and the MLMA association statistic is based on univariate GLS and is

$$\hat{\beta}_{j,MLMA} = \frac{\mathbf{X}_j^\top \mathbf{V}^{-1} \mathbf{y}}{\mathbf{X}_j^\top \mathbf{V}^{-1} \mathbf{X}_j}$$

where $\mathbf{V}$ is the phenotypic variance covariance matrix:

$$\mathbf{V} = Var[\mathbf{y}] = \sigma_g^2 \mathbf{A} + \sigma_e^2 \mathbf{I} = \frac{h^2}{M}(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)$$

Strictly speaking, the $\mathbf{V}$ matrix should be based on a GRM which includes all SNPs except the SNP for which the association statistic is calculated, so that this SNP is not fitted twice in the model. This is impractical because it would require to compute a different GRM for

each SNP. What is therefore often done in practice is to have 23 GRMs, where each chromosome is left out in one of them, and each SNP is modelled with the GRM that doesn't include that SNP (GCTA MLMA-LOCO).

When we compare the definition of $\hat{\beta}_{j,MLMA}$ here to $\hat{\beta}_{j,BLUP}$, we see that they are closely related. In fact, $\hat{\beta}_{j,MLMA}$ can be expressed like this:

$$\hat{\beta}_{j,MLMA} = \frac{M}{h^2} \frac{\hat{\beta}_{j,BLUP}}{\mathbf{X}_j^\top \mathbf{V}^{-1} \mathbf{X}_j}$$

The un-standardized MLMA effects can be obtained like this:

$$\hat{\beta^*}_{j,MLMA} = \frac{\mathbf{X}_j^{*\top} \mathbf{V}^{-1} \mathbf{y}}{\mathbf{X}_j^{*\top} \mathbf{V}^{-1} \mathbf{X}_j^*} \approx \frac{\hat{\beta}_{j,MLMA}}{\sqrt{(2p_j(1-p_j))}}$$

```
V = h2*grm + (1-h2)*diag(n)
Vi = solve(V)
beta_mlma = (t(x) %*% Vi %*% y) / diag(t(x) %*% Vi %*% x)
beta01_mlma = beta_mlma * sqrt(varx)

p1 = qplot(beta_gwas, beta_mlma, col=ldscores) + geom_abline() + theme()
+
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta_gwas, beta_mlma), 3))) +
  scale_colour_continuous(low='black', high='green')
p2 = qplot(beta_blup, beta_mlma, col=ldscores) + geom_abline() + theme()
+
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta_blup, beta_mlma), 3))) +
  scale_colour_continuous(low='black', high='green')
grid.arrange(p1, p2, ncol=2)
```

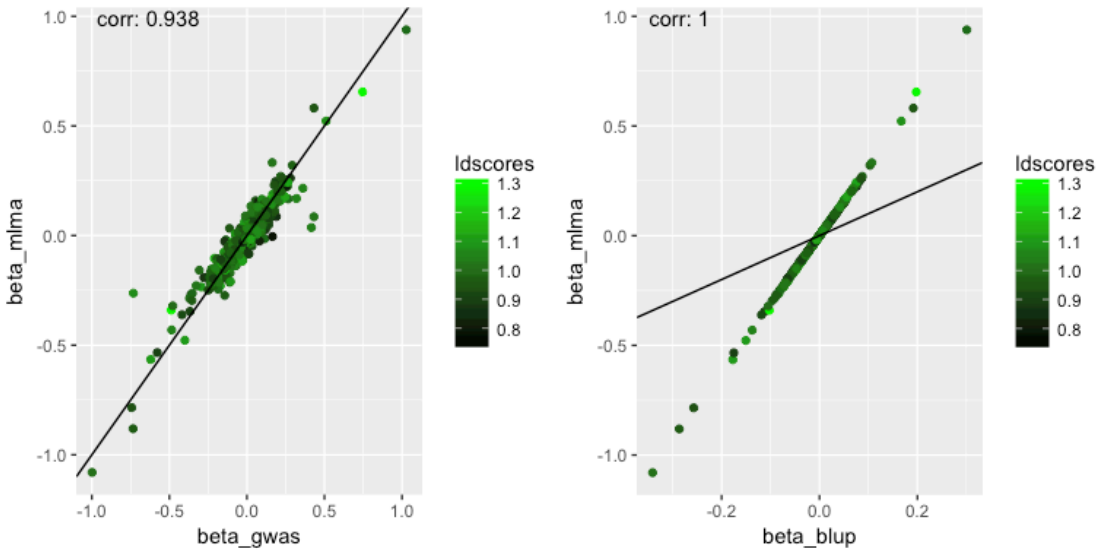

Figure 53: MLMA estimates compared to other estimates in data without LD

MLMA estimates are not shrunk and therefore about as large as GWAS estimates.

Let's see what this comparison looks like in data with a wider range in LD scores.

```
Vld = h2*grmld + (1-h2)*diag(n)
Vild = solve(Vld)
beta_mlmald = (t(x01ld) %*% Vild %*% yld) / diag(t(x01ld) %*% Vild %*%
x01ld) / sqrt(varxld)
beta01_mlmald = beta_mlmald * sqrt(varxld)

p1 = qplot(beta_gwasld, beta_mlmald, col=ldscoresld) + geom_abline() +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta_gwasld, beta_mlmald),
3))) +
  scale_colour_continuous(low='black', high='green')
p2 = qplot(beta_blupld, beta_mlmald, col=ldscoresld) +
  geom_point() + geom_abline() + theme() +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=paste0('corr: ', round(cor(beta_blupld, beta_mlmald),
3))) +
  scale_colour_continuous(low='black', high='green')
grid.arrange(p1, p2, ncol=2)
```

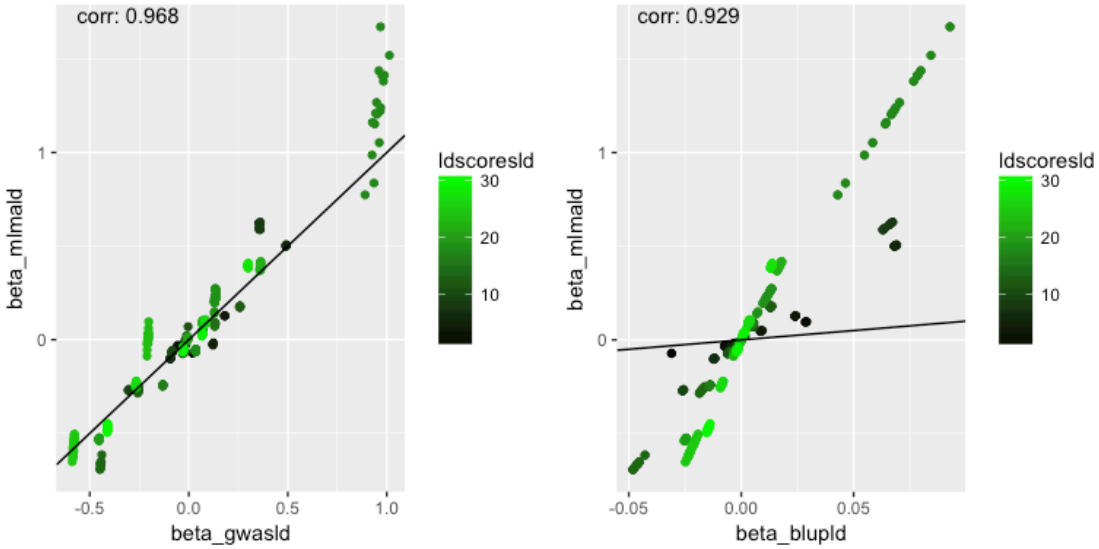

Figure 54: MLMA estimates compared to other estimates in data with LD

In data with more variable LD between SNPs, MLMA estimates are more similar to the marginal GWAS estimates than to the conditional BLUP estimates. Conditioned on LD score, BLUP estimates are proportional to MLMA estimates.

Comparison of the models

$$\hat{\beta}_{j,GWAS} = \frac{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{y}}{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{X}_j}$$

The covariance of SNP and phenotype, scaled by the variance of the SNP.

$$\hat{\beta}_{j,OLS} = \frac{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{y}}{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{X}_j}$$

The covariance of SNP and phenotype, scaled by the covariance structure of all SNPs.

$$\hat{\beta}_{j,BLUP} = \frac{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{y}}{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{X}_j}$$

The covariance of SNP and phenotype, scaled by the covariance structure of all SNPs, without overemphasizing the covariances among SNPs.

$$\hat{\beta}_{j,MLMA} = \frac{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{y}}{(\lambda_M \mathbf{I} + \mathbf{X}^\top \mathbf{X})_j^{-1} \mathbf{X}_j^\top \mathbf{X}_j}$$

Like $\hat{\beta}_{j,BLUP}$, but scaled back so that effects are not shrunk if they are in LD with many other SNPs.

SNP effects - precision of estimates

Almost as important as the estimate itself is the standard error of the estimate. This quantifies the variability of the estimate that is due to the sampling process and thus the precision of the estimate. The standard error of an estimate is the basis for other statistics like p-values and confidence intervals and is defined as

$$SE^2_{\hat{\beta}} = var(\hat{\beta} \mid \beta) = var(\hat{\beta} - \beta)$$

We don't usually know β, but we can still estimate what the standard error of an effect estimate is.

Sampling variance of GWAS estimates

The standard error of the GWAS estimate of the effect size of a SNP is:

$$SE^2_{\hat{\beta^*}_{j,GWAS}} = var(\hat{\beta^*}_{j,GWAS} \mid \beta^*_j) = \frac{var(\mathbf{y}) - var(\mathbf{X}^*_j)\,\hat{\beta^*}^2_{j,GWAS}}{N \times var(\mathbf{X}^*_j)}$$

Here the numerator represents the variance in the phenotype that is not explained by this SNP. Since under a polygenic model, each SNP explains almost no variance ($var(\mathbf{X}^*_j)\,\hat{\beta^*}^2_{j,GWAS}$ is close to 0), and $var(\mathbf{y})$ is often one, the numerator is often approximated as one:

$$SE^2_{\hat{\beta^*}_{j,GWAS}} = var(\hat{\beta^*}_{j,GWAS} \mid \beta^*_j) \approx \frac{1}{N \times var(\mathbf{X}^*_j)} = \frac{1}{N \times 2p_j(1-p_j)}$$

It gets even simpler when switching to standardized genotypes and effect sizes: Since $var(\mathbf{X}_j) = 1$ for every SNP j, the sampling variance of $\hat{\beta}_{j,GWAS}$ is

$$SE^2_{\hat{\beta}_{j,GWAS}} = var(\hat{\beta}_{j,GWAS} \mid \beta) \approx \frac{1}{N}$$

Let's see how the expected sampling variance compares to the observed sampling variance, stratified by MAF:

```
beta_se = sqrt(1/(n*varx))
observed_beta_sampling_variance = tapply(beta_gwas - beta, bins, var)
expected_beta_sampling_variance = tapply(beta_se^2, bins, mean)

dat = data.frame(binNum=1:length(binMeans),
                 expected_beta_sampling_variance,
                 observed_beta_sampling_variance)
ggplot(dat, aes(expected_beta_sampling_variance,
observed_beta_sampling_variance)) +
  geom_text(aes(label=binNum, col=binMeans)) +
  scale_x_log10() + scale_y_log10() +
  geom_abline() +
  xlab(expression(paste(hat(SE)[hat(beta)]^2,
                        '  (',hat(var),'(',hat(beta),' | ',beta,'))'))) +
  ylab(expression(paste('var(',hat(beta),' | ',beta,')')))
```

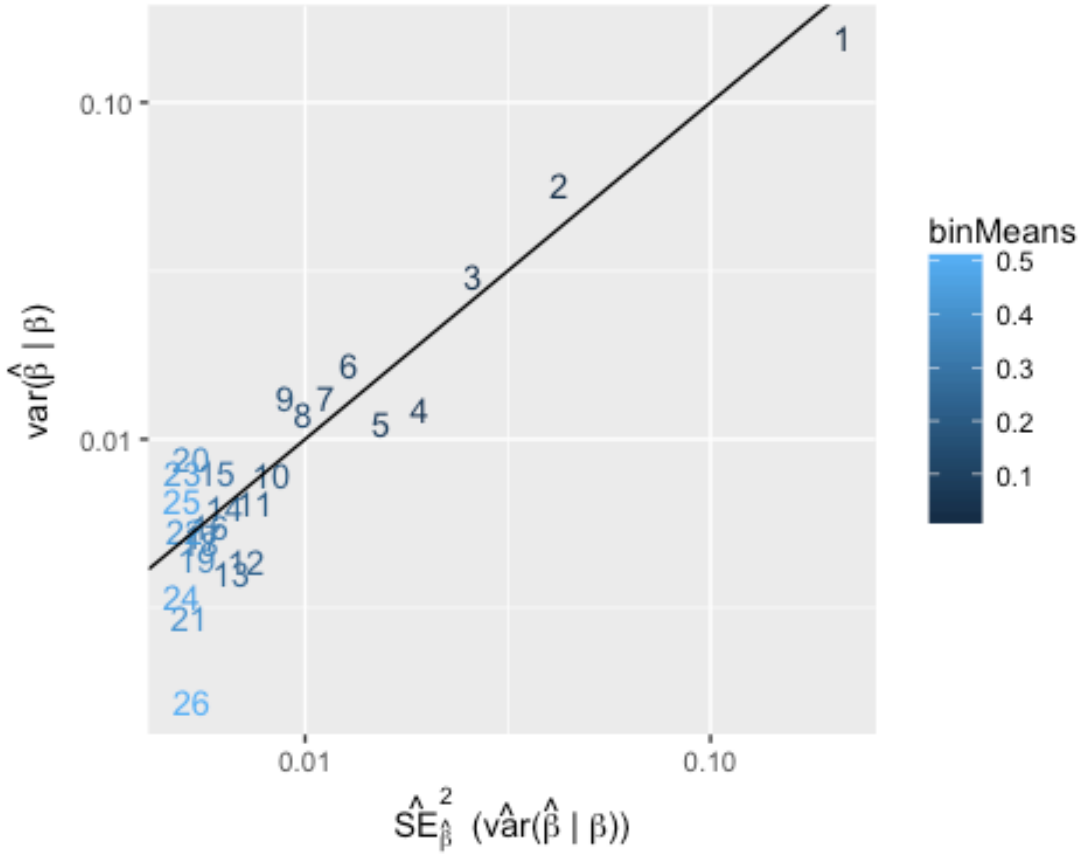

Figure 55: The standard error of $\hat{\beta}_{GWAS}$ is higher for rare SNPs

Sampling variance of MLMA estimates

If $var(\mathbf{y} = 1)$ and if each SNP explains only very little variance, the sampling variance of $\hat{\beta^*}_{j,GWAS}$ can be approximated as

$$SE^2_{\hat{\beta^*}_{j,GWAS}} \approx \frac{1}{N \times var(\mathbf{X}_j^*)} = \frac{1}{\mathbf{X}_j^{*\top}\mathbf{X}_j^*}$$

Similarly, the standard error for MLMA estimates can be approximated as

$$SE^2_{\hat{\beta^*}_{j,MLMA}} \approx \frac{1}{\mathbf{X}_j^{*\top}\mathbf{V}^{-1}\mathbf{X}_j^*}$$

z-scores

z-scores are defined as

$$z = \frac{x - \mathbb{E}[x]}{\sigma(x)}$$

where $\sigma(x)$ is the standard deviation of x. It quantifies how far a statistic is from its expectation in standard deviation units and it should follow a standard normal distribution. We can calculate the z-score of an effect estimate under the null hypothesis ($\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0$)

$$z_j = \frac{\hat{\beta^*}_j}{SE_{\hat{\beta^*}_j}}$$

Since the χ^2_k distribution with k degrees of freedom is the same as a sum of k independent standard normal distributions, z-scores are closely related to χ^2 values with degree of freedom of 1:

$$z^2 = \chi^2_1$$

Strictly speaking, $\frac{\hat{\beta^*}_j}{SE_{\hat{\beta^*}_j}}$ follows a t-distribution with $N - 1$ degrees of freedom, not a standard normal distribution, but for large N the difference becomes negligible.

p-value of GWAS estimates

The z-score tells us where on the null distribution our effect estimate lies. It can therefore be converted into a (two-sided) p-value, which is the probability of observing an effect at least as extreme as the one estimated, under the null distribution.

$$\begin{aligned} \text{p} - \text{value}_{\hat{\beta^*}_j} &= 2 \times min\{Pr(\hat{\beta^*} \geq \hat{\beta^*}_j \mid \beta^* = 0), Pr(\hat{\beta^*} \leq \hat{\beta^*}_j \mid \beta^* = 0)\} \\ &= 2\Phi(-\frac{|\hat{\beta^*}_j|}{SE_{\hat{\beta^*_j}}}) \\ &= 2\Phi(-|\hat{z}|) \end{aligned}$$

where $\Phi(x)$ is the probability distribution function of the standard normal distribution at x (pnorm) and $\hat{\beta^*}$ denotes the random variable rather than the concrete estimate of $\hat{\beta^*}_j$.

To go from two-sided p-values to z-scores:

$$z = \Phi^{-1}(\frac{\text{p} - \text{value}}{2})$$

where $\Phi^{-1}(x)$ is the quantile function of the standard normal distribution at x (qnorm).

```
lim = 3
z = 1.5
ggplot(data.frame(x=c(-lim, lim)), aes(x)) + stat_function(fun=dnorm) +
  stat_function(fun=dnorm, geom='area', xlim=c(z, lim), fill='purple',
alpha = 0.2) +
  stat_function(fun=dnorm, geom='area', xlim=c(-lim, -z), fill='purple',
alpha = 0.2) +
  geom_vline(xintercept=z, linetype=2) +
  annotate('text', z+.11, .2,

label=as.character(expression(paste("z=",frac(hat(beta),SE[beta])))),
           parse=T, hjust=0) +
  ylab(expression(phi(x)))
```

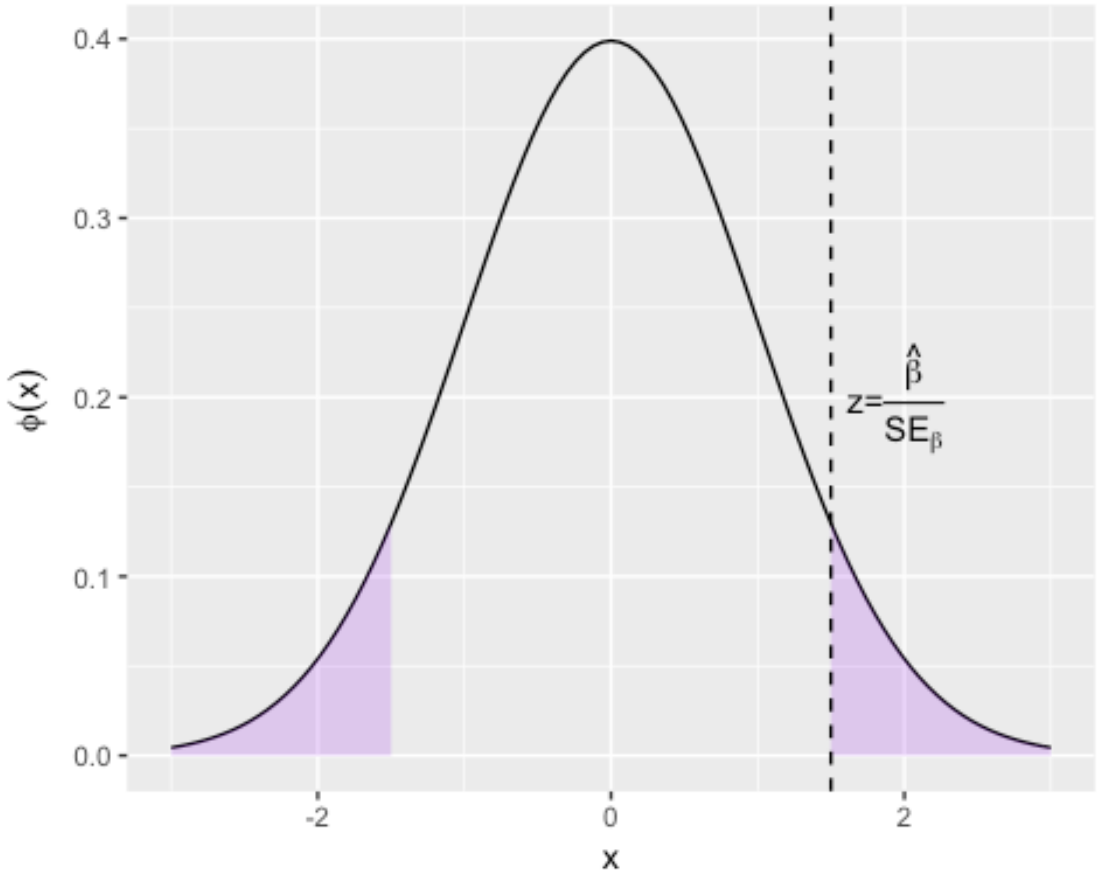

Figure 56: z-score and p-value

The vertical line represents a z-score. The total area of the shaded segments represents the p-value.

```
par(mfrow=c(1,4))
z = beta_gwas/beta_se
pval = 2*pnorm(-abs(z))

plot(beta_gwas, -log10(pval), main='Volcano plot')
hist(pval, 100, col='black')
qqPlot(pval, main='qq-plot')
plot(-log10(pval), main='Manhattan plot')
```

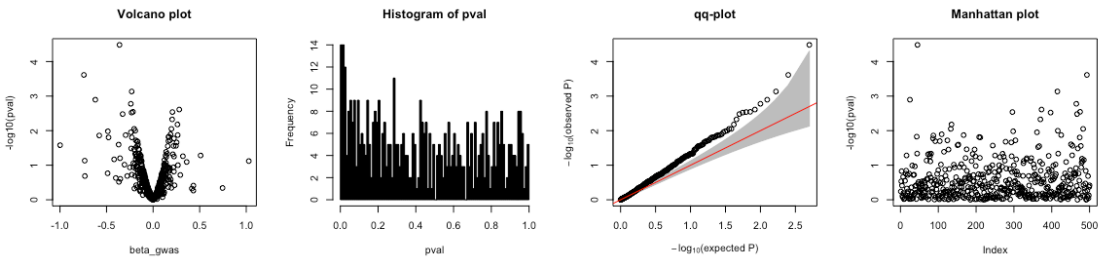

Figure 57: Different ways to visualize p-values.

Most are based on negative log_{10} transformed p-values, to better highlight the more interesting small p-values. Volcano plots illustrate that SNPs with larger absolute $\hat{\boldsymbol{\beta}}^*$ values tend to have lower p-values. Histograms of p-values are useful, because under the null hypothesis, p-values will follow a uniform distribution between 0 and 1. QQ-plots can also highlight an overall inflation of p-values, but with better resolution at the lower end. Manhattan plots from GWAS p-values usually show peaks of multiple SNPs with low p-values, because of LD.

```
par(mfrow=c(1,1))
```

Re-estimating GWAS statistics

The equations above can be rearranged to estimate various statistics. To keep things simpler, in this section SE^2 refers to $SE^2_{\hat{\beta}^*_{j,GWAS}}$ and β refers to $\hat{\beta}^*_j$. The approximations assume that $2p(1-p)\beta^2$ is close to zero.

$$SE^2 = \frac{var(\mathbf{y}) - 2p(1-p)\beta^2}{N \times 2p(1-p)} \approx \frac{var(\mathbf{y})}{N \times 2p(1-p)}$$

$$N = \frac{var(\mathbf{y}) - 2p(1-p)\beta^2}{SE^2 \times 2p(1-p)} \approx \frac{var(\mathbf{y})}{SE^2 \times 2p(1-p)}$$

$$var(\mathbf{y}) = 2p(1-p)(N \times SE^2 + \beta^2) \approx 2p(1-p) \times N \times SE^2$$

$$\beta^2 = \frac{var(\mathbf{y})}{2p(1-p)} - N \times SE^2$$

This one doesn't tell us about the sign of β, however.

$$p(1-p) = \frac{var(\mathbf{y})}{2(N \times SE^2 + \beta^2)}$$

The last one can be solved for p, and only leaves two options which center around 0.5.

When working with GWAS summary statistics, all of these quantities except $var(\mathbf{y})$ are usually known. When $var(\mathbf{y})$ can be assumed to be one, sample size or standard error can be re-estimated from the other quantities. When β refers to standardized β, $2p(1-p)$ should be set to 1.

Confidence intervals and power

While p-values are widely used, they often draw criticism for their potential to be misused. An alternative to reporting p-values can be to report confidence intervals, which describe the range in which the estimated value would fall with a certain probability, if the sampling process was repeated. If the sampling distribution is normal, the 95% confidence interval can be estimated as $\hat{\boldsymbol{\beta}} \pm 1.96 \times SE_{\boldsymbol{\beta}}$ (because $1 - 2\Phi(-1.96) \approx 0.95$). A 95% confidence interval that does not include 0 is equivalent to a p-value smaller than 0.05 for an effect being different from 0.

```
plot_power(n=200, reps=30, b=0.2, xlim=c(-.25,.5), ylim=c(0,6))
```

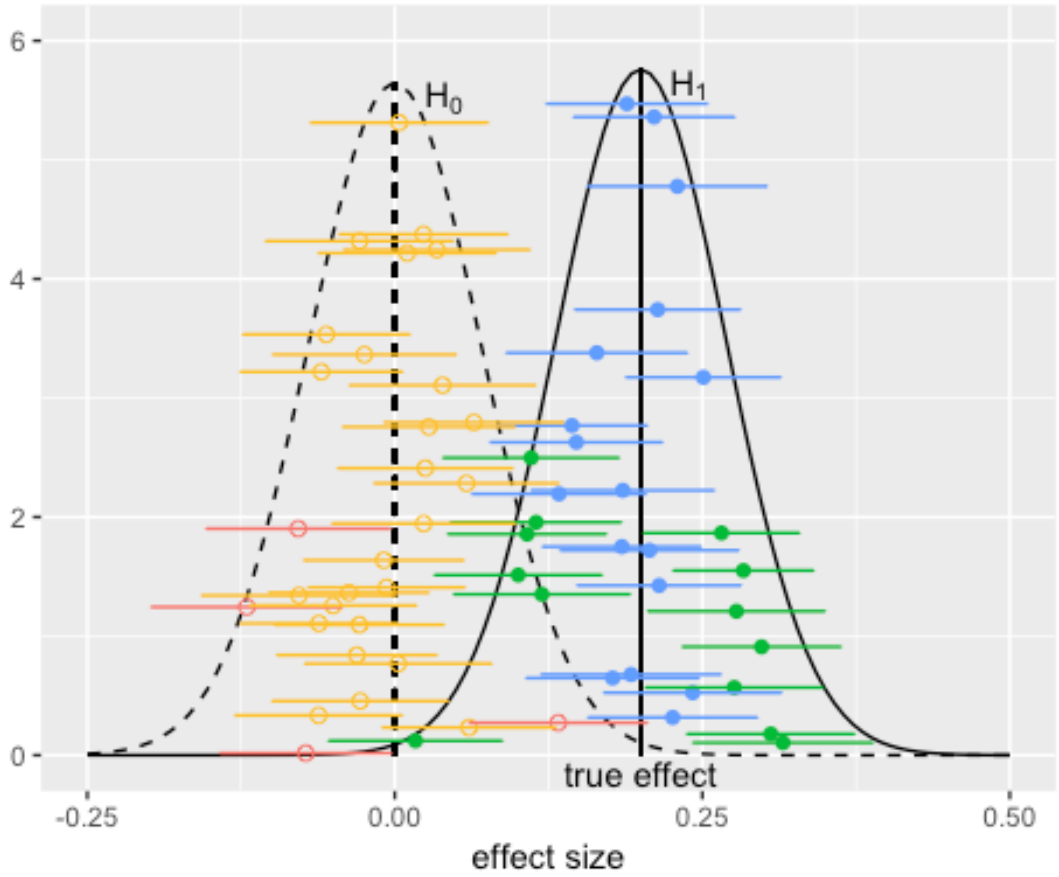

Figure 58: Draws from the sampling distribution and from the null distribution

This plot shows a number of draws from the null distribution and the same number of draws from the sampling distribution of the effect estimate. The horizontal bars are standard errors, and are like a 68% confidence interval (because $1-2\Phi(-1)\approx 0.68$). Each standard error represents an estimate of what the (square root of the) variance of the sampling distribution is.

An important question in hypothesis testing is this: What is the probability that we will be able to reject the null hypothesis ($\hat{\boldsymbol{\beta}}=0$), if there is a true effect of a certain magnitude? This probability depends at least on the sample size and on how willing we are to mistake a null effect for a true effect (type I error, α). Not rejecting the null hypothesis when there is a true effect is called a type II error and is the opposite of power. This means that being strict in controlling the type I error rate (small α) will reduce power.

```
plot_power(n=200, reps=30, b=0.2, alpha=0.05, show_points=F, xlim=c(-
.25,.5), ylim=c(0,6))
```

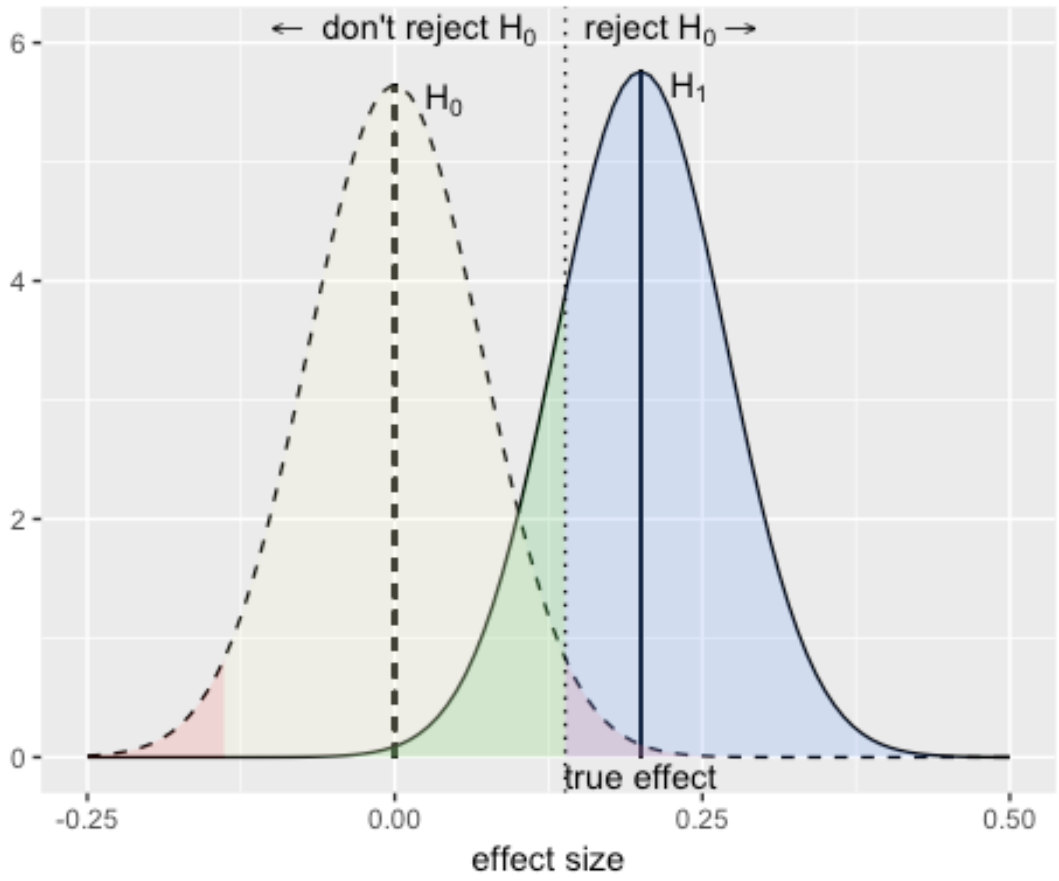

Figure 59: Power visualized

In this plot α is the red area (as a proportion of the area under the null distribution (H_0)), and the power is the blue area (as a proportion of the area under the true effect sampling distribution (H_1)). Power is also $1 - \beta$, where β is the type II error rate (green area).

```
plot_power(n=40, reps=30, b=0.2, alpha=0.05, show_points=F, xlim=c(-
.25,.5), ylim=c(0,6))
```

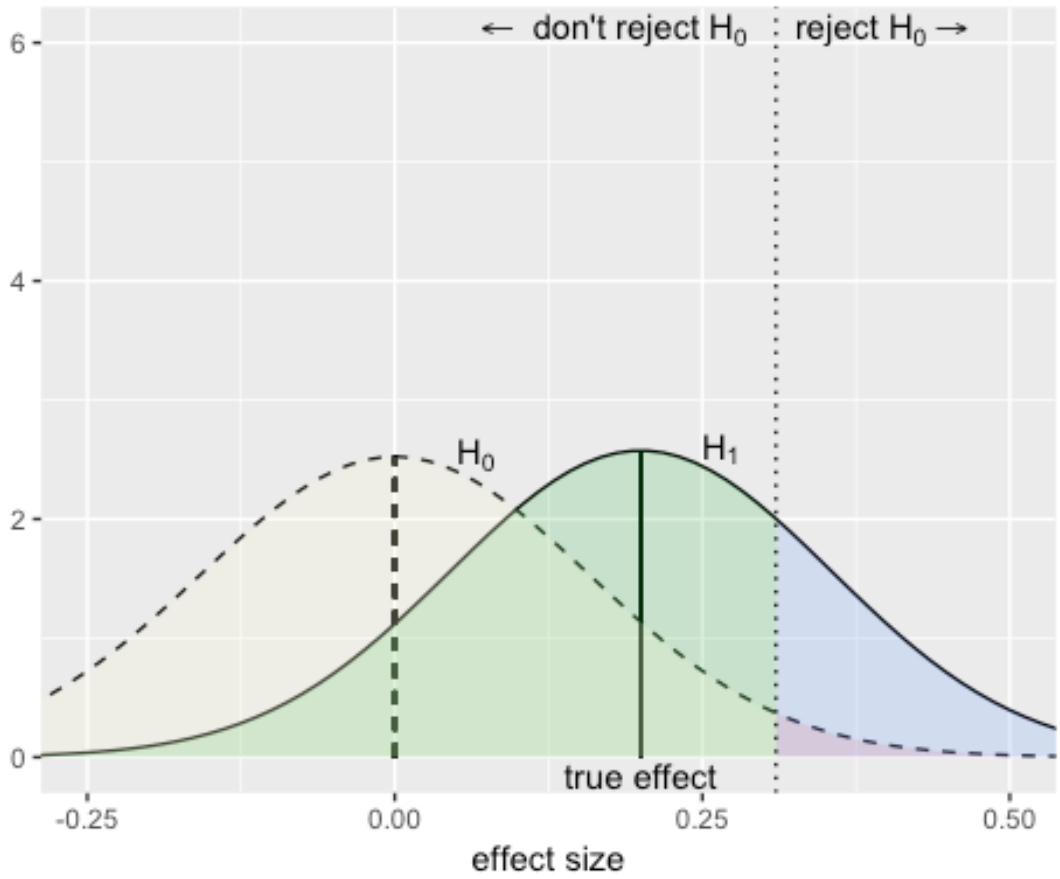

Figure 60: The effect of sample size on power

Lowering sample size will widen both distributions, and thereby increases the threshold at which an estimated effect can be considered significant at the same level of α, thus lowering power.

```
plot_power(n=40, reps=100, b=0.2, alpha=0.05, show_points=T, show_se=F,
           xlim=c(-.25,.5), ylim=c(0,6))
```

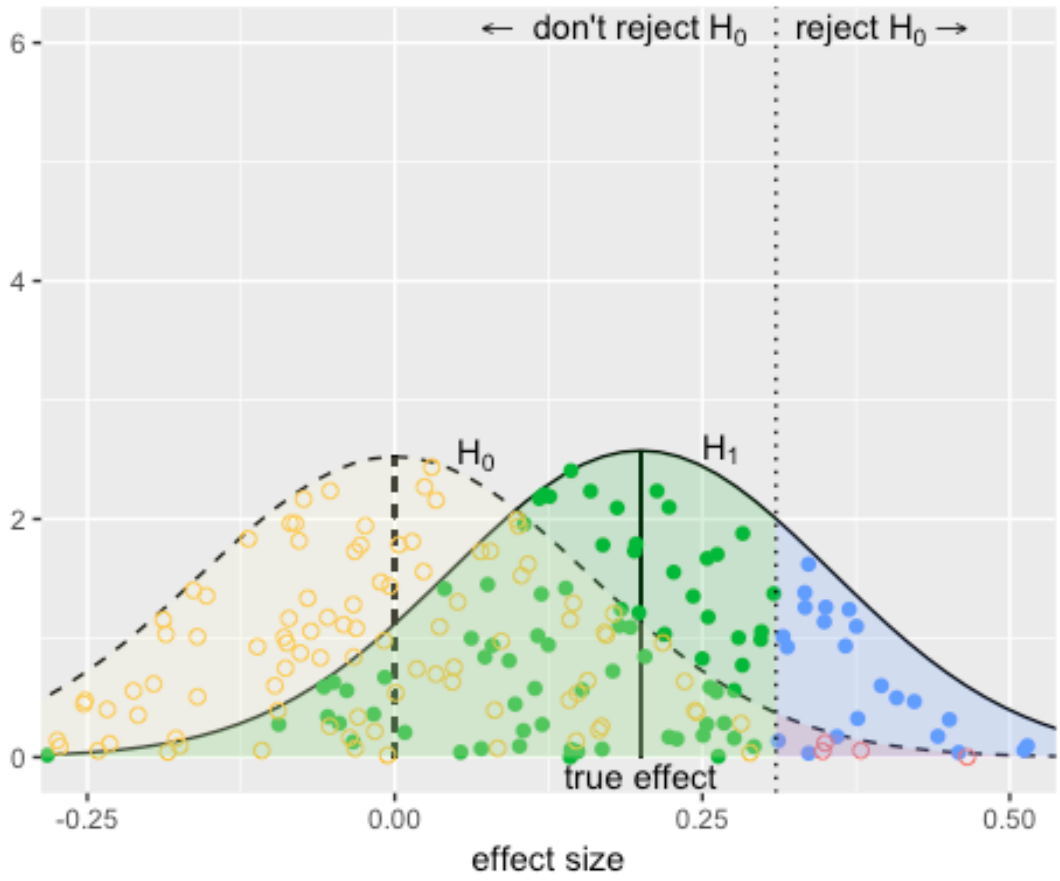

Figure 61: Power determines the proportion of true positive results

Evaluating in which of the four quadrants estimated effects fall, allows to estimate the numbers of true and false positives and negatives, and many statistics derived from these values, such as the area under the ROC curve (AUC).

Here, the null distribution and the sampling distribution of the true effect both follow normal distributions (t-distributions actually, but we ignore that). When the null distribution is a χ^2 distribution instead, the true effect distribution will follow a non-central χ^2 distribution.

Prediction error variance (PEV) of BLUP estimates

The precision of BLUP estimates is usually given as a variance-covariance matrix, rather than as a scalar number and is usually defined for the genetic effects rather than the SNP effects, which is why it is called prediction error variance, but here we define it for the SNP effects:

$$Var\left[\hat{\boldsymbol{\beta}}_{BLUP} \mid \boldsymbol{\beta}\right] = Var\left[\hat{\boldsymbol{\beta}}_{BLUP} - \boldsymbol{\beta}\right]$$

$$= Var[\boldsymbol{\beta}] + Var\left[\hat{\boldsymbol{\beta}}_{BLUP}\right] - 2Cov\left[\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{BLUP}\right]$$

$$= Var[\boldsymbol{\beta}] + Var\left[\hat{\boldsymbol{\beta}}_{BLUP}\right] - 2Var\left[\hat{\boldsymbol{\beta}}_{BLUP}\right]$$

$$= Var[\boldsymbol{\beta}] - Var[\hat{\boldsymbol{\beta}}_{BLUP}]$$

Notice that while $var(\hat{\boldsymbol{\beta}}_{GWAS}) > var(\boldsymbol{\beta})$, $var(\hat{\boldsymbol{\beta}}_{BLUP}) < var(\boldsymbol{\beta})$.

From before, we know that $Var[\boldsymbol{\beta}] = \frac{h^2}{M}\mathbf{I}_M$

Recall that the BLUP solution can be written as:

$$\hat{\boldsymbol{\beta}}_{BLUP} = \frac{h^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$$

The variance of that is:

$$Var[\hat{\boldsymbol{\beta}}_{BLUP}] = Var[\frac{h^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}] = (\frac{h^2}{M})^2\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}$$

Therefore,

$$Var[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{BLUP}] = Var[\boldsymbol{\beta}] - Var[\hat{\boldsymbol{\beta}}_{BLUP}] = \frac{h^2}{M}(\mathbf{I}_M - \frac{h^2}{M}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})$$

The diagonal elements of this matrix should be similar to $var(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{BLUP})$.

```
beta01_blup_se2 = var(beta01_blup - beta01)
pev = (h2/m) * (diag(m) - (h2/m) * t(x01) %*% Vi %*% x01)

c(beta01_blup_se2, mean(diag(pev)))
## [1] 0.0006994066 0.0006773369
```

However, this is not as useful as the standard error estimates given above, because it is a function of the data, not just of the parameters. The section on the expected accuracy of a BLUP predictor has an expression which serves at the same time as an estimate of the BLUP PEV as a function of the parameters:

$$cor^2(\mathbf{y}, \hat{\mathbf{g}}) = var(\hat{\mathbf{g}}) = R^2 = \frac{h^2}{1 + \frac{M_e(1-R^2)}{Nh^2}}$$

$$var(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{BLUP}) = var(\boldsymbol{\beta}) - var(\hat{\boldsymbol{\beta}}_{BLUP}) = \frac{h^2}{M} - \frac{R^2}{M} = \frac{h^2 - R^2}{M}$$

```
R2_BLUP = function(m, n, h2) {
  k = m/n
  ( (k + h2) - sqrt( (k+h2)^2 - 4*k*h2^2) ) / (2*k)
}

c(beta01_blup_se2, mean(diag(pev)), (h2 - R2_BLUP(m, n, h2))/m)
## [1] 0.0006994066 0.0006773369 0.0006770330
```

The BLUP PEV is substantially lower than the standard errors of the GWAS effect estimate. This is only partially due to the higher accuracy of BLUP estimates and mostly reflects the downward bias of BLUP estimates.

Winner's curse and unbiasedness

When the SNP with the largest effect is being selected from a GWAS, its estimated effect size is likely going to be larger than its true effect size. This phenomenon is called winner's curse. If we estimated the effect of the same SNP again in and independent cohort, we would likely get a smaller estimate in the new cohort. This closely related effect is called regression to the mean and is just a consequence of how we selected the SNP and the imperfect correlation between the true and the estimated effect size.

If we estimated the effect for this same SNP many times in independent cohorts, the effect estimates will center around the SNP's true effect size. GWAS estimates are therefore unbiased in the sense that conditioned on the true effect, the estimated effect will in expectation be the same as the true effect:

$$\mathbb{E}[\hat{\beta} \mid \beta] = \beta$$

However, when we select the SNP with the largest estimated effect size, we don't condition on the true effect, we condition on the estimated effect. So we know the estimated effect, and ideally we would like an estimate that tells us what the true effect size is, in expectation. It's therefore often handy to have another definition of unbiasedness, which says that the estimated effect is equal to the expectation of the true effect, conditioned on the estimated effect:

$$\mathbb{E}[\beta \mid \hat{\beta}] = \hat{\beta}$$

GWAS estimates (OLS, which is a BLUE, best linear unbiased estimators) are unbiased in the first sense, but biased in the second (they suffer from winner's curse but are not shrunk), whereas BLUP (best linear unbiased prediction) estimates are biased in the first sense but unbiased in the second (they don't suffer from winner's curse but are shrunk).

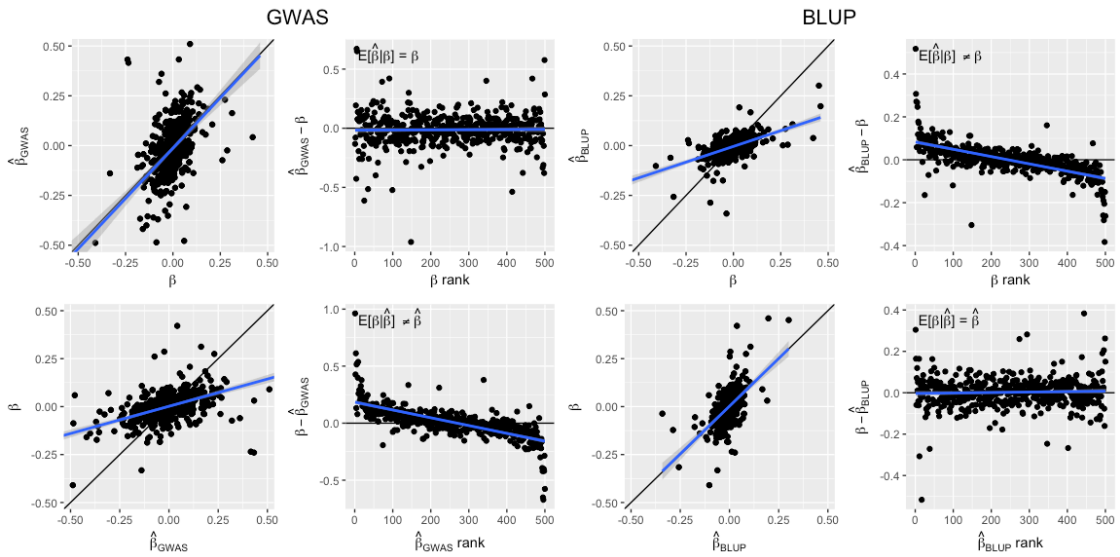

Figure 62: Two definitions of unbiasedness visualized

First row: conditioning on true effects. Second row: conditioning on estimated effects.

Both meanings of unbiasedness have an implication on what the covariance between true and estimated effect sizes is. For unbiasedness in the first (OLS) sense:

$$cov(\beta, \hat{\beta}_{GWAS}) = var(\beta)$$

and for unbiasedness in the second (BLUP) sense:

$$cov(\beta, \hat{\beta}_{BLUP}) = var(\hat{\beta}_{BLUP})$$

This is why the regression slopes in the above plots are around one, if the unbiasedness condition is met ($\text{slope of y} \quad \text{x} = \frac{cov(x,y)}{var(x)}$).

Fixed effects vs random effects

By now the terms fixed effects and random effects have been mentioned a few times without having been defined. If you, as I have, looked online what the definition of fixed and random effects is, chances are you are a bit confused at this point. There are many different definitions of these terms that have little in common with one another. One thing that

everyone can agree on is that a mixed effects model gets its name from having both fixed and random effects.

Here is how the terms fixed and random effects are used in the present context: SNP effects are modeled as random effects, everything else (age, sex, a couple of principal components, cohort) is modeled as fixed effects (here, we don't have any fixed effects for simplicity). In concrete terms that means that in a mixed effects model OLS or GLS (generalized least squares) is used to estimate fixed effects (without shrinkage) and BLUP is used to estimate random effects (with shrinkage). Both can be estimated simultaneously in a mixed effects model to account for the covariance between fixed and random effects (see Henderson's mixed model equation). Usually the fixed effects are much smaller in number than the random effect (SNPs). That means that there is a practical reason for modelling SNPs as random: With $M > N$ it is impossible to estimate the effect of each SNP without shrinkage.

Sometimes you will also hear that random effects are modelled as coming from a random distribution, whereas fixed effects are not. This is equivalent to the above statement about shrinking effect sizes, because BLUP can be viewed as a Bayesian method which assumes that the SNP effects come from a normal prior distribution with a mean of 0 and a variance that is determined by h^2: Low h^2 means the prior distribution will have a small variance, so all effects will be shrunk heavily towards 0. A large h^2 means that the prior distribution will have a large variance, and so it will be less informative and will have a smaller effect on the posterior distribution of effect sizes (the BLUP estimates), so that the effects will be shrunk less. A fixed effect is thus equivalent to a random effect with a prior distribution that has infinite variance and will not be shrunk at all. Growing sample size will not affect the absolute value of the shrinkage factor λ, but it will make it smaller relative to the diagonal elements in $\mathbf{X}^\top\mathbf{X}$, so the prior distribution will have less effect on the posterior distribution.

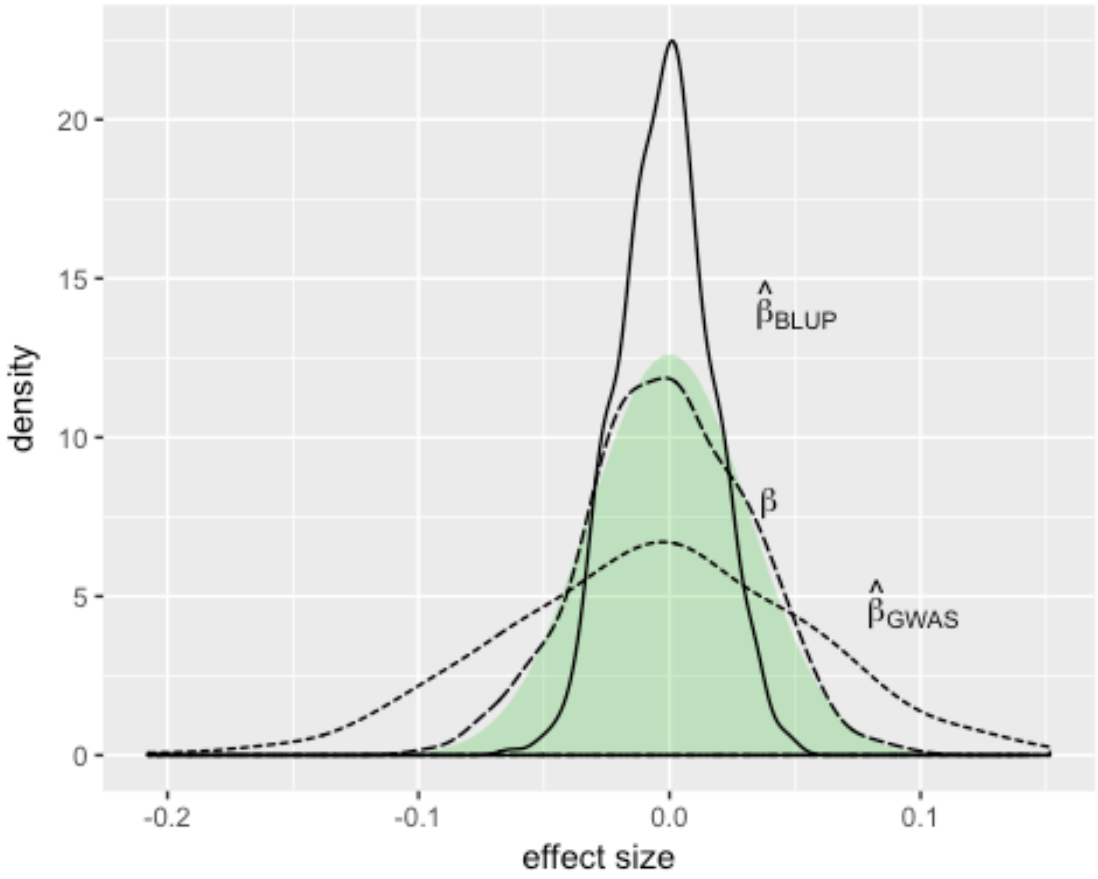

Figure 63: Distribution of true and estimated effect sizes

The green shaded area represents the prior distribution for the BLUP effect size estimation and should closely follow the distribution of the true effect sizes. The "prior distribution" of the GWAS effect estimates has infinite variance.

This is in contrast with another definition of random effects, which occurs in the context of hierarchical models with multiple observations in different groupings, for example multiple observations per individual. According to this definition, the effect of an individual would be treated as a random, if the people on which data have been collected were of no particular interest, and if those people are rather seen as random sample of a bigger population.

The SNP model is an adaptation from the earlier pedigree model which was used to estimate breeding values of cows and other animals. This pedigree model does indeed have a hierarchical, nested structure in which multiple cows belonging to the same family are being grouped together. A random effect in the pedigree model is a random effect according to both definitions. With the introduction of SNP genotyping arrays, the model has been adapted so that relationships between cows were not modeled anymore by grouping them

into families, but by treating them all as unrelated and calculating their genetic similarity using genome wide SNP data. The pedigree based relatedness matrix has been replaced by the GRM, and the term random effects for breeding values / genetic effects is still used, but the term random effects now only refers to the first definition.

Meta-analysis

Often GWAS are performed on smaller cohorts. The summary statistics ($\hat{\beta}_{GWAS}$ and $SE_{\hat{\beta}_{GWAS}}$) of multiple smaller GWAS can be combined to increase power. This is in contrast to a Mega-analysis in which individual level genotype data from multiple small cohorts is combined.

In this and the next section, $\hat{\beta}$ is short for $\hat{\beta}_{j,GWAS}$ and can also refer to the non-standardized version of the effect estimate ($\hat{\beta}^*_{j,GWAS}$).

Meta analyzed $\hat{\beta}$ values over c cohorts can be obtained as:

$$\hat{\beta}_{META} = \frac{\sum^c \frac{\hat{\beta}_c}{SE^2_{\hat{\beta}_c}}}{\sum^c \frac{1}{SE^2_{\hat{\beta}_c}}} \qquad SE_{\hat{\beta}_{META}} = \sqrt{\frac{1}{\sum^c \frac{1}{SE^2_{\hat{\beta}_c}}}}$$

Let's get meta-analyzed summary statistics first.

```
cohort_stats = list()
num = 10
first = floor(seq(1, n+1, len=num+1))
last = first[-1]-1
for(i in 1:num) {
  nc = last[i] - first[i]
  xc = x[first[i]:last[i],]
```

```
  yc = y[first[i]:last[i],]
  betac = t(xc) %*% yc / diag(t(xc)%*%xc)
  sec = sqrt(1/(varx*nc))
  cohort_stats[[i]] = data.frame(betac, sec)
}
```

And now meta-analyze them, one cohort at a time.

```
numer_one_cohort = lapply(cohort_stats, function(x) x$betac/x$sec^2)
denom_one_cohort = lapply(cohort_stats, function(x) 1/x$sec^2)
numer_sums = do.call('cbind', Reduce(`+`, numer_one_cohort, accumulate =
TRUE))
denom_sums = do.call('cbind', Reduce(`+`, denom_one_cohort, accumulate =
TRUE))
beta_meta = numer_sums/denom_sums

dat = data.frame(maf, beta_gwas, beta_meta=beta_meta[,num])
ggplot(dat, aes(beta_gwas, beta_meta, col=log10(maf))) +
  geom_point() + geom_abline()
```

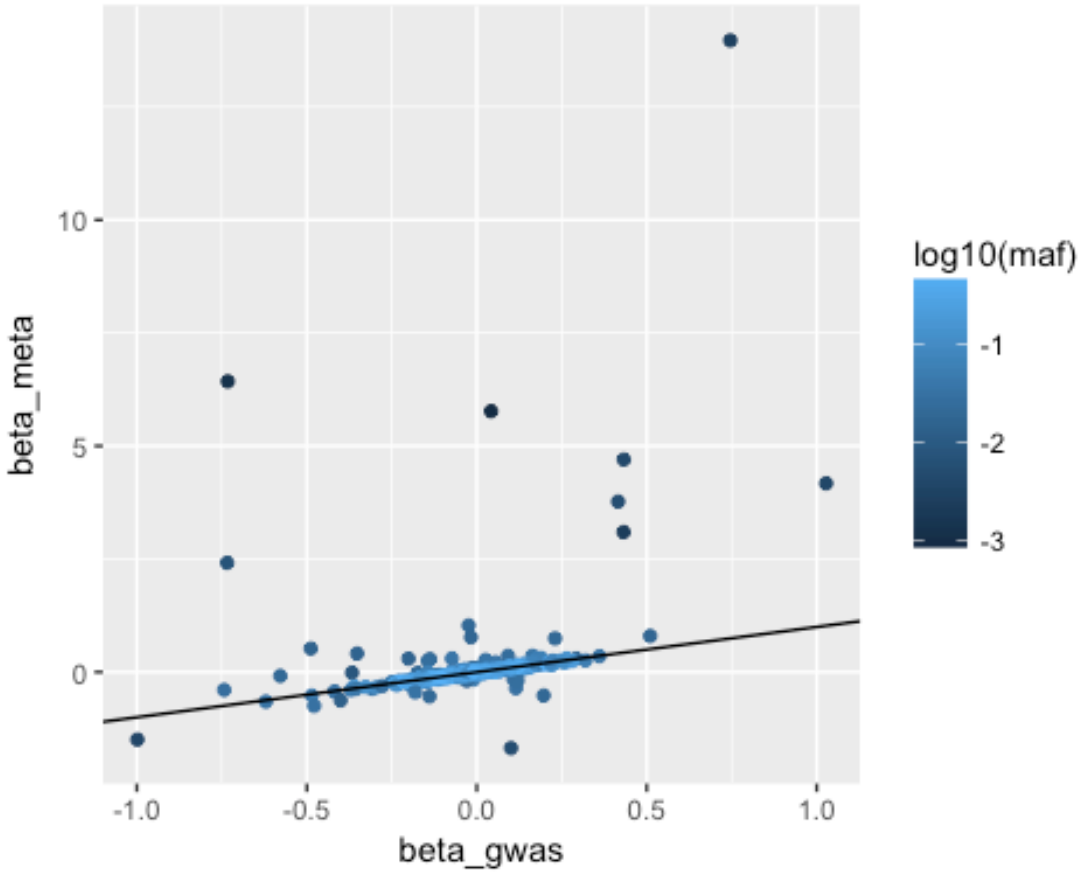

Figure 64: Effect estimates from a meta-analysis and a mega-analysis

Low MAF SNPs are more likely to be outliers.

```
minmaf = .1
cor_mega = cor(beta[maf > minmaf], beta_gwas[maf > minmaf])
corrs = sapply(1:num, function(i) cor(beta[maf > minmaf], beta_meta[maf >
minmaf, i]))

dat = data.frame(name=c(num+1, 1:num), corr=c(cor_mega, corrs))
ggplot(dat, aes(as.factor(name), corr, fill=corr)) +
  geom_col() +
  scale_x_discrete(labels=c(1:num, 'mega')) +
  xlab('number of cohorts in meta-analysis') +
  ylab('correlation between meta-analyzed and true effects') +
  scale_fill_gradient(low='black', high='purple')
```

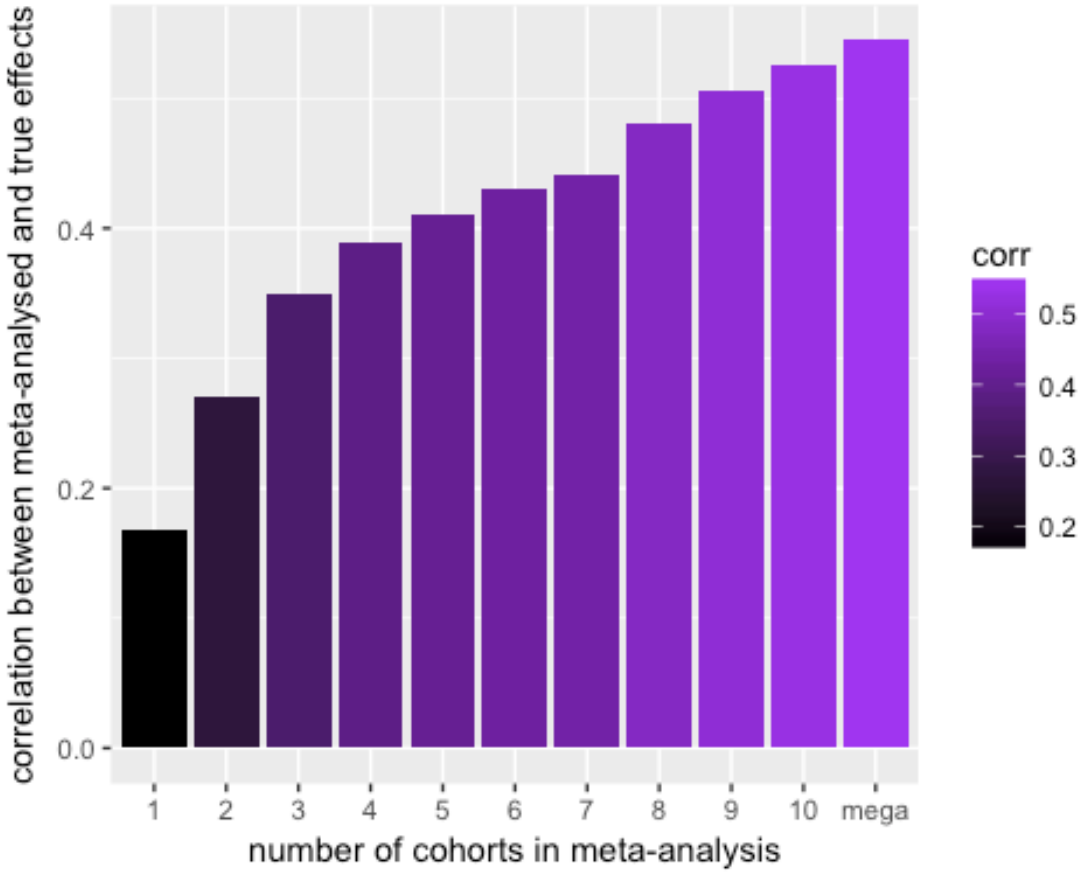

Figure 65: Sequentially adding more cohorts to a meta-analysis

De-meta-analysis

It is also possible to exclude the effects of one cohort from GWAS summary statistics, using a de-meta-analysis:

$$\hat{\beta}_{DE-META} = \hat{\beta}_{META} - SE^2_{\hat{\beta}_{DE-META}} \sum^{c} \frac{\hat{\beta}_c - \hat{\beta}_{META}}{SE^2_{\hat{\beta}_c}}$$

$$SE_{\hat{\beta}_{DE-META}} = \sqrt{\frac{1}{\frac{1}{SE^2_{\hat{\beta}_{META}}} - \sum^{c} \frac{1}{SE^2_{\hat{\beta}_c}}}}$$

Variance of GWAS estimates

We have previously looked at the sampling variance of a SNP effect estimate, $var(\hat{\beta}_{j,GWAS} \mid \beta_j)$. Often variance of a SNP effect estimate, $var(\hat{\beta}_{j,GWAS})$, is also of interest. While $var(\hat{\beta}_{j,GWAS} \mid \beta_j)$ estimates the variance of $\hat{\beta}_{j,GWAS}$, assuming that the true effect size is kept constant but the sampling process is repeated, $var(\hat{\beta}_{j,GWAS})$ estimates the variance of $\hat{\beta}_{j,GWAS}$ assuming that both the assignment of a true effect and the sampling process are repeated.

If $var(\hat{\beta}_{j,GWAS})$ is i.i.d. for all SNPs j, $var(\hat{\beta}_{j,GWAS})$ will be equal to the variance across all SNPs, $var(\hat{\boldsymbol{\beta}}_{GWAS})$. The true effect sizes are i.i.d., and this is why $var(\beta_j) = var(\boldsymbol{\beta}) = \frac{h^2}{M}$.

Variance of standardized GWAS estimates

From an earlier section we know that if $var(\mathbf{y}) = 1$ and the variance explained per SNP is small,

$$var(\hat{\beta}_{j,GWAS} \mid \beta_j) = var(\hat{\beta}_{j,GWAS} - \beta_j) = \frac{1}{N}$$

We also know that

$$var(\beta_j) = \frac{h^2}{M}$$

From this we are able to get $var(\hat{\beta}_{j,GWAS})$.

$var(X - Y) = var(X) + var(Y) - 2 \times cov(X, Y)$, so:

$$var(\hat{\beta}_j - \beta_j) = var(\hat{\beta}_j) + var(\beta_j) - 2 \times cov(\hat{\beta}_j, \beta_j)$$

OLS estimates are unbiased, in the sense that $\mathbb{E}[\hat{\beta}_j \mid \beta_j] = \beta_j$. This implies that $cov(\beta_j, \hat{\beta}_j) = var(\beta_j)$.

From this it follows that

$$var(\hat{\boldsymbol{\beta}}_{GWAS}) = var(\hat{\beta}_{j,GWAS}) = var(\beta_j) + var(\hat{\beta}_{j,GWAS} - \beta_j) = \frac{h^2}{M} + \frac{1}{N}$$

The effect of LD on the variance of GWAS estimates

The derivation above uses the unbiasedness property of OLS estimates. However, the unbiasedness and consistency of OLS estimates depends on certain assumptions, which are not all met in a GWAS setting. Specifically, the model is not correctly specified, if the model doesn't include important explanatory factors for $\mathbf{y}$ that are correlated with $\mathbf{X}_j$ (SNPs in LD with SNP j). Not including these other SNPs can lead to omitted-variable bias. It's easy to see that this will create bias when you consider only two SNPs which are highly correlated and together explain most of the variance. Including only one of them in the model will induce a correlation between the error term and $\mathbf{X}_j$ and lead to a biased estimate, where the bias depends on the LD between the two SNPs.

In a polygenic setting, this bias will not be very large, but can still be noticeable, especially if LD is widespread and N is large compared to M. In this case, a better approximation of $var(\hat{\beta}_{j,GWAS})$ is

$$var(\hat{\beta}_{j,GWAS}) = l_j \frac{h^2}{M} + \frac{1}{N}$$

where l_j is the LD-score of SNP j.

The LD score regression method is based on this observation.

Averaged over all SNPs, this becomes

$$var(\hat{\boldsymbol{\beta}}_{GWAS}) = \frac{h^2}{M_e} + \frac{1}{N}$$

```
dat = data.frame(type1=c('observed', 'observed', 'expected', 'expected',
'observed', 'expected'),
                 type2=c('true_effects', 'estimated_effects',
'true_effects',
                         'estimated_effects', 'estimated_effects_ld',
'estimated_effects_ld'),
```

```
                value=c(var(beta01), var(beta01_gwas), h2/m, h2/m + 1/n,
                        var(beta01_gwasld), h2/meld + 1/n  ))
dat$type2 = factor(dat$type2, levels=levels(factor(dat$type2))[c(3,1,2)])

ggplot(dat, aes(type1, value, fill=type1)) +
  geom_col() +
  facet_wrap(~ type2, scales='free_x') +
  xlab('') +
  ylab(expression(paste("var(",hat(beta),") / var(",beta,")"))) +
  theme(legend.position = 'none')
```

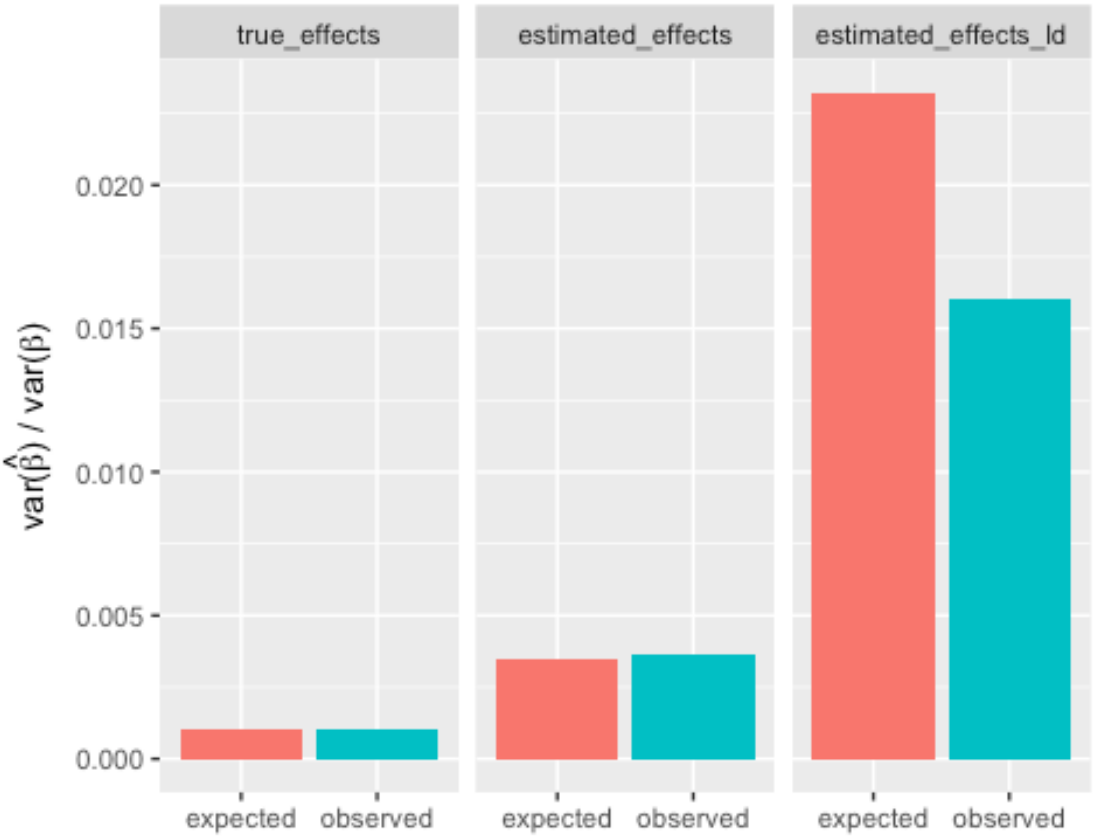

Figure 66: Expected and observed beta variance.

For the genotypes without LD, the variance across all estimates is well approximated by $\frac{h^2}{M} + \frac{1}{N}$*, for the genotypes with LD, the variance across all estimates is better approximated by* $\frac{h^2}{M_e} + \frac{1}{N}$

Variance of un-standardized GWAS estimates

$var(\hat{\beta}_{j,GWAS})$ is identically distributed for each SNP for each SNP, if β_j is identically distributed for each SNP. But the same is not true for $\hat{\beta}^*_{j,GWAS}$, which is $\hat{\beta}^*_{j,GWAS} = \frac{\hat{\beta}_{j,GWAS}}{\sqrt{2p_j(1-p_j)}}$. Here, the variance of rare SNPs is larger than that of common SNPs. A SNP with MAF zero has infinite variance, and this makes the expectation of $var(\hat{\boldsymbol{\beta}}^*_{GWAS})$ infinte too, for a MAF range beginning at zero.

Comparison of different BLUP formulations

The BLUP equation here looks very different from more common formulations of the BLUP model, for example in the animal breeding literature. This is partly due to different notation and partly because the animal model is a pedigree model in which different individuals are modeled as being part of different families. In the model that is used here, there is no such hierarchical structure, which simplifies the model. Lynch and Walsh (Eq. 26.4) defines BLUP (individual effects) as follows:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^\top\mathbf{V}^{-1}(y - \mathbf{X}\hat{\boldsymbol{\beta}})$$

This equation uses a different notation and makes different assumptions than what is presented in this document. The following table compares the notation.

Table 20: Comparison of BLUP models

Lynch and Walsh (pedigree model)	Meaning	This document (SNP model)
$\hat{\mathbf{u}}$	estimated genetic effects	$\hat{\mathbf{g}}$
$\mathbf{G}$	covariance matrix of random genetic effects: pedigree matrix or SNP based matrix	$\sigma_g \mathbf{A}$
$\mathbf{Z}$	Incidence matrix in case of pedigree design. For unrelated individuals this is a diagonal matrix and can be ignored	not relevant ($\mathbf{I}$)
$\mathbf{V} = \mathbf{ZGZ}^\top + \mathbf{R}$	variance-covariance matrix of phenotype vector	$\mathbf{V} = \sigma_g \mathbf{A} + \sigma_e \mathbf{I}_N$
y	phenotypes. In L&W, not standardized. Here, it is assumed that they have mean 0, variance 1, so $\sigma_g = h^2$. Further, here we assume that y represents residuals after accounting for covariates or fixed effects.	$\mathbf{y}$
$\mathbf{X}$	Incidence matrix of fixed effects	not relevant (0)
$\hat{\boldsymbol{\beta}}$	Effect sizes of fixed effects	not relevant (0)

After substituting L&W terminology with the terminology used here, the equation above can be rewritten as

From this, it follows that

$$\hat{\boldsymbol{\beta}}_{BLUP} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_M)^{-1} \mathbf{X}^\top \mathbf{y}$$

and so the two models are identical if we assume the absence of fixed effects.

Prediction

Estimated SNP effects ($\hat{\boldsymbol{\beta}}$) can easily be converted into predictors of genetic effects for individuals ($\hat{\mathbf{g}}$):

$$\hat{g}_i = \sum^{j} X_{ij}\,\hat{\beta}_j$$

In matrix notation:

$$\hat{\mathbf{g}} = \mathbf{X}\,\hat{\boldsymbol{\beta}}$$

Unless we have a prediction model that incorporates non-genetic factors, this is also our phenotype predictor:

$$\hat{\mathbf{y}} = \hat{\mathbf{g}} = \mathbf{X}\,\hat{\boldsymbol{\beta}}$$

Generally, more accurate $\hat{\boldsymbol{\beta}}$ will result in more accurate $\hat{\mathbf{g}}$.

The importance of an independent test set

The first rule of any kind of prediction is that the training set, which is used to train a predictor, has to be strictly separated from the testing set, which is used to evaluate the predictor. Otherwise the accuracy of the predictor will be highly overestimated.

That is why we will first create an independent test set of equal size and equal allele frequencies, and then use the true effect sizes to again simulate phenotypes in this independent set.

```
x012test = apply(x012, 2, sample)
xtest = scale(x012test, scale=FALSE)
x01test = scale(x012test, scale=TRUE)


# simulate test set phenotypes
gtest = xtest %*% beta
etest = rnorm(n, 0, sqrt(1-h2))
ytest = gtest + etest
```

```
# same for data with LD
x012ldtest = jitter(x012test[,rep(1:m, 1:m)[1:m]], .03)
x01ldtest = scale(x012ldtest)

g01ldtest = x01ldtest %*% beta
eldtest = rnorm(n, 0, sqrt(1-h2))
yldtest = g01ldtest + eldtest
```

Creating individual predictors from SNP effects is very simple. Apart from scaling, the following is equivalent to `plink --score`.

```
ghat_train_gwas = x01 %*% beta01_gwas
ghat_test_gwas = x01test %*% beta01_gwas

# for data with LD
ghat_test_gwasld = x01ldtest %*% beta01_gwasld
# confounded data set
x01conf = rbind(x01, x01test[1:floor(n/10),])
yconf = c(y, ytest[1:floor(n/10)])
beta01_gwas_conf = t(x01conf) %*% yconf / diag(t(x01conf)%*%x01conf)

ghat_conf_gwas = x01test %*% beta01_gwas_conf

ghat_train_blup = x01 %*% beta01_blup
ghat_test_blup = x01test %*% beta01_blup

dat = data.frame(predictor=c('training', 'test', 'test_confounded_10%'),
                 corr=c(cor(y, ghat_train_gwas),
                        cor(ytest, ghat_test_gwas),
                        cor(ytest, ghat_conf_gwas)))
ggplot(dat, aes(predictor, corr, fill=predictor)) +
  geom_col() +
```

```
  xlab('') + ylab('correlation') +
  theme(legend.position='none')
```

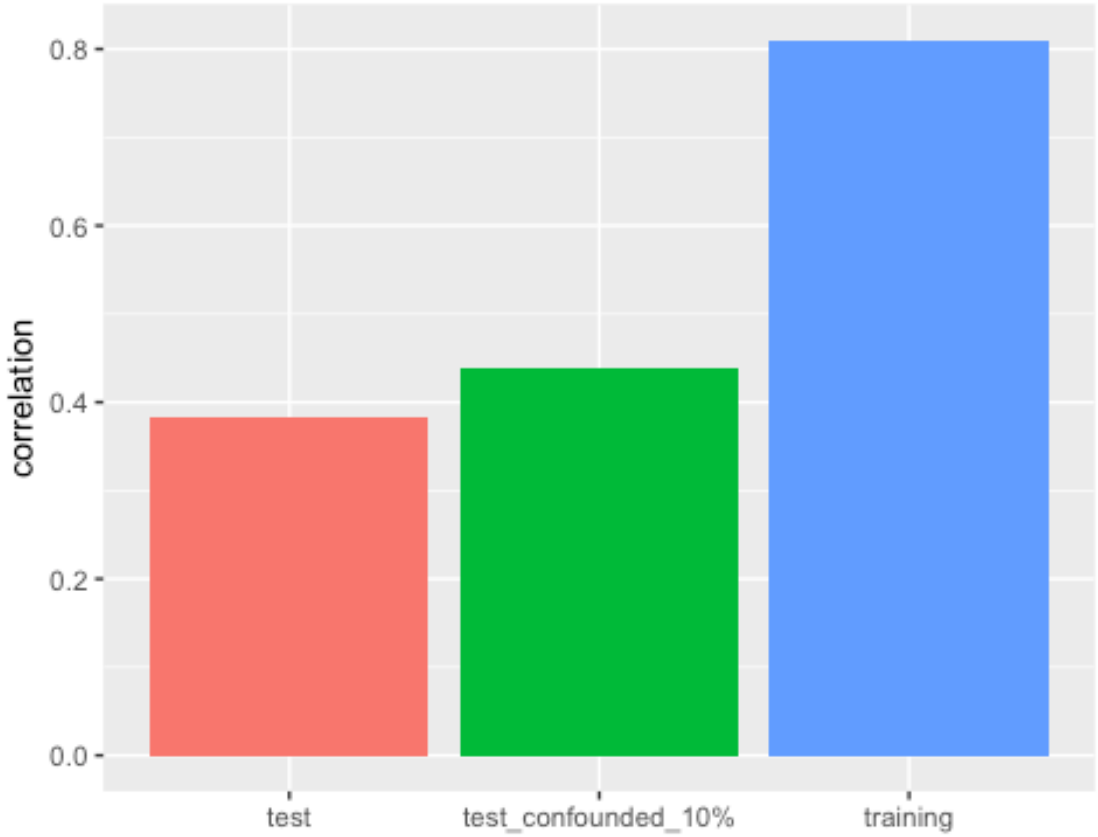

Figure 67: Prediction accuracy if training and test set are not independent

The red bar represents prediction accuracy in an independent set. The green bar is prediction accuracy in an independent set, but 10% of the independent set went into the training set, so it's not independent anymore. The blue bar represent accuracy in the same sample that was used to estimate SNP effects. Only the red bar is a useful measure of prediction accuracy.

Accuracy of a GWAS predictor

The expected accuracy of a GWAS predictor, measured as the correlation between predicted breeding value and the phenotype in an independent sample, is given by:

$$cor^2(\mathbf{y}, \hat{\mathbf{g}}_{GWAS}) = \frac{h^2}{1 + \frac{M_e}{Nh^2}}$$

In case you wonder why:

$$\begin{aligned}
cor^2(\mathbf{y}, \hat{\mathbf{g}}_{GWAS}) &= \frac{cov^2(\mathbf{y}, \hat{\mathbf{g}}_{GWAS})}{var(\mathbf{y})var(\hat{\mathbf{g}}_{GWAS})} && \text{definition of correlation} \\
&= \frac{cov^2(\mathbf{g} + \mathbf{e}, \hat{\mathbf{g}}_{GWAS})}{var(\hat{\mathbf{g}}_{GWAS})} && var(\mathbf{y}) = 1, \text{definition of } \mathbf{y} \\
&= \frac{cov^2(\sum^j \mathbf{X}_j \beta_j, \sum^j \mathbf{X}_j \hat{\beta}_{j,GWAS})}{var(\sum^j \mathbf{X}_j \hat{\beta}_{j,GWAS})} && \mathbf{g} \text{ and } \mathbf{e} \text{ are uncorrelated, definition of } \mathbf{g},\ \hat{\mathbf{g}}_{GWAS} \\
&= \frac{M^2 \times cov^2(\beta_j, \hat{\beta}_{j,GWAS})}{M \times var(\hat{\beta}_{j,GWAS})} && \text{assuming SNPs are independent} \\
&= \frac{M \times cov^2(\beta, \hat{\beta}_{GWAS})}{var(\hat{\beta}_{GWAS})} && \beta \text{ are i. i. d.} \\
&= \frac{M \times var^2(\beta)}{var(\hat{\beta}_{GWAS})} && \text{follows from OLS unbiasedness} \\
&= \frac{M(\frac{h^2}{M})^2}{\frac{h^2}{M} + \frac{1}{N}} && \text{derived in earlier sections} \\
&= \frac{h^2}{1 + \frac{M}{Nh^2}} && \text{rearrange}
\end{aligned}$$

This derivation assumes independent markers, which means that $M = M_e$. In order for the equation to be applicable to data sets with non-independent markers, M has to be replaced by M_e.

Accuracy of a BLUP predictor

In the derivation above, we used an estimate of $var(\hat{\beta}_{GWAS})$ of $\frac{h2}{M_e}+\frac{1}{N}$, where $\frac{1}{N}$ is the standard error of the estimate. In a model that fits all SNPs at the same time, the standard error is better estimated as $\frac{1-R^2}{N}$, where R^2 is the variance explained by all SNP effect estimates. This is at the same time the expected accuracy of our predictor, and is smaller than an estimate of the variance explained by all SNPs (SNP-heritability), because each SNP has an estimation error. If we use this estimate of the standard error, then the expected prediction accuracy becomes:

$$cor^2(\mathbf{y},\hat{\mathbf{g}}) = R^2 = \frac{h^2}{1+\frac{M_e(1-R^2)}{Nh^2}}$$

This doesn't account for the overfitting problem that $\hat{\beta}_{OLS}$ estimates have, so it is not a good estimate of the accuracy of a predictor based on $\hat{\beta}_{OLS}$, but it is a good estimator of the accuracy of a $\hat{\beta}_{BLUP}$ based predictor.

R^2 occurs on both sides of the equation, but we can solve for R^2:

$$cor^2(\mathbf{y},\hat{\mathbf{g}}_{BLUP}) = R^2 = \frac{k+h^2-\sqrt{(k+h^2)^2-4kh^2}}{2k}$$

where $k=\frac{M_e}{N}$

Note that $cor^2(\mathbf{g},\hat{\mathbf{g}})$ will always be larger than $cor^2(\mathbf{y},\hat{\mathbf{g}})$ by a factor of $\frac{1}{h^2}$, because $\mathbf{g}$ doesn't contain the error component of $\mathbf{y}$:

$$cor^2(\mathbf{g},\hat{\mathbf{g}}_{BLUP}) = \frac{1}{1+\frac{M_e(1-R^2)}{Nh^2}}$$

```
R2_GWAS = function(m, n, h2) {
  h2 / (1 + (m / (n*h2)))
}

R2_BLUP = function(m, n, h2) {
```

```
  k = m/n
  ( (k + h2) - sqrt( (k+h2)^2 - 4*k*h2^2) ) / (2*k)
}
cor.test.plus = function(x, y, ...) {
  # like cor.test, but also returns se of correlation
  corr = cor.test(x, y, ...)
  corr$se = unname(sqrt((1 - corr$estimate^2)/corr$parameter))
  corr
}

ghat_test_mlma = x01test %*% beta01_mlma
ghat_test_mlmald = x01ldtest %*% beta01_mlmald
ghat_test_blup = x01test %*% beta01_blup
ghat_test_blupld = x01ldtest %*% beta01_blupld

cornold = t(sapply(list(ghat_test_gwas, ghat_test_mlma, ghat_test_blup),
         function(x) {corr = cor.test.plus(ytest, x); c(corr$estimate,
corr$se)}))
corld = t(sapply(list(ghat_test_gwasld, ghat_test_mlmald,
ghat_test_blupld),
         function(x) {corr = cor.test.plus(yldtest, x); c(corr$estimate,
corr$se)}))
dat = data.frame(type=rep(c('GWAS', 'MLMA', 'BLUP'), 2),
                 ld = c(rep(' without LD', 3), rep('with LD', 3)),
                 n=n,
                 corr=c(cornold[,1], corld[,1]),
                 se=c(cornold[,2], corld[,2]))
dat$type = factor(dat$type, levels=levels(factor(dat$type))[c(2,3,1)])

nlim = c(ceiling(n/10000), n*300)
mypal = rev(c('seagreen4', 'royalblue4'))
ggplot(dat[dat$type != 'MLMA',], aes(n, corr, col=type)) +
  stat_function(fun=function(...) sqrt(R2_GWAS(...)), args=list(h2=h2,
m=me), col=mypal[1]) +
```

```
  stat_function(fun=function(...) sqrt(R2_BLUP(...)), args=list(h2=h2,
m=me), col=mypal[2]) +
  stat_function(fun=function(...) sqrt(R2_GWAS(...)), args=list(h2=h2,
m=meld), col=mypal[1], linetype=2) +
  stat_function(fun=function(...) sqrt(R2_BLUP(...)), args=list(h2=h2,
m=meld), col=mypal[2], linetype=2) +
  scale_x_log10(limits=c(nlim[1], nlim[2])) +
  geom_point(size=2) +
  geom_errorbar(aes(ymin=corr-se, ymax=corr+se, linetype=ld), width=.1) +
  scale_colour_manual(values=mypal) + xlab('sample size') +
ylab('correlation') +
  geom_hline(yintercept = sqrt(h2), linetype=2)
```

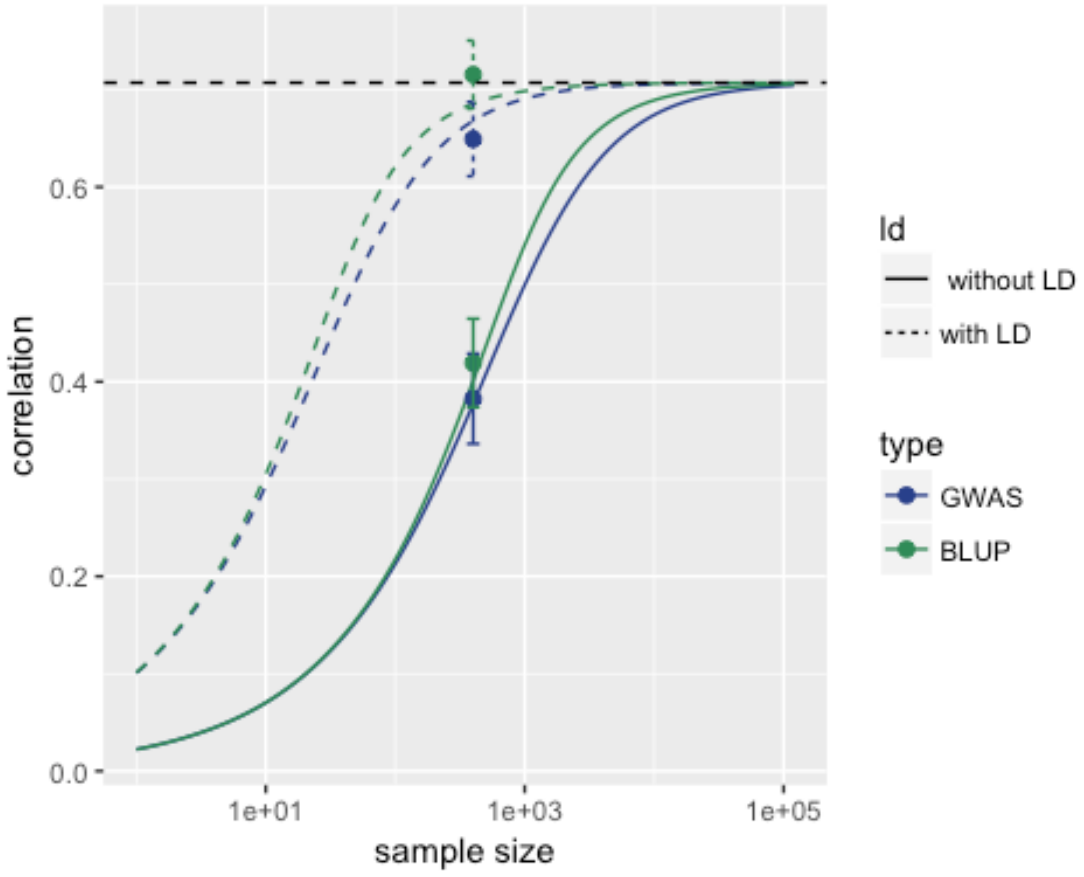

Figure 68: Expected and observed prediction accuracy

Prediction accuracy +/- SE of the GWAS and BLUP predictor. Lines show expected accuracy for a range of N. The horizontal line represents $\sqrt{h^2}$, which is the upper limit for an (additive) genetic predictor. Dashed lines show accuracies for data with LD. β estimates are not more accurate in data without LD, but the combined effect estimate of each LD block is

more accurate, since it is averaged over many SNPs. This is why prediction accuracy is higher in data with LD, if the LD structure is the same in the test set. The standard errors of the point estimates are underestimated for the simulations with LD, because the SE estimate assumes that all individual predictors are independent. However, they are not independent because they depend on correlated β estimates.

From the unbiasedness of BLUP predictors, it follows that R^2 is at the same time the expected variance of a BLUP predictor:

$$R^2 = cor^2(\mathbf{y}, \hat{\mathbf{g}}_{BLUP}) = \frac{cov^2(\mathbf{y}, \hat{\mathbf{g}}_{BLUP})}{var(\hat{\mathbf{g}}_{BLUP})} = var(\hat{\mathbf{g}}_{BLUP})$$

SNP selection

The predictors so far have been based on all SNPs. In practice people often select SNPs based on their LD with other SNPs or based on their p-value.

The motivation behind selecting SNPs based on LD with other SNPs is that GWAS estimates are marginal effect estimates and so regions with high LD will influence the predictor too much.

The motivation behind selecting SNPs based on p-value, is that SNPs with low p-value are more likely to be truly associated with the trait, and so the signal to noise ratio of the SNP effects that go into the predictor will increase. This is especially true in well powered studies, where the p-value better separates the causal SNPs from the rest.

Since we simulate SNP effects from a normal distribution, all SNPs will have some effect, but even here, selecting based on p-value can slightly increase accuracy.

```
ghatmat = t(t(xtest) * beta_gwas[,1])
ghatmat_pval = ghatmat[,order(pval)]
ghatmat_cumsum = t(apply(ghatmat_pval, 1, cumsum))
corrs = sapply(1:m, function(i) cor(ytest[,1], ghatmat_cumsum[,i]))


reps = 5
```

```
corrs_rand = replicate(reps, {
  ghatmat_pval = ghatmat[,sample(1:m)]
  ghatmat_cumsum = t(apply(ghatmat_pval, 1, cumsum))
  sapply(1:m, function(i) cor(ytest[,1], ghatmat_cumsum[,i]))
})

dat = rbind(data.frame(snp=1:m, rep=reps+1, corr=corrs),
            melt(corrs_rand) %>% rename(snp=Var1, rep=Var2, corr=value))
maxx = which.max(dat$corr[dat$rep==reps+1])
maxcor = dat[maxx, 'corr']
bestsnp = dat[maxx, 'snp']
p_opt = sort(pval)[bestsnp]

ggplot(dat, aes(snp, corr, col=as.factor(rep))) +
  geom_line() +
  geom_hline(yintercept=maxcor, linetype=2) +
  geom_vline(xintercept=bestsnp, linetype=3) +
  xlab('number of SNPs included in predictor') +
  ylab('correlation') +
  scale_color_manual(values=c(rep('grey', reps), 'navy')) +
  theme(legend.position='none') +
  annotate('text', bestsnp, 0, label=paste0('p-value = ', round(p_opt,
2)), hjust=1.1)
```

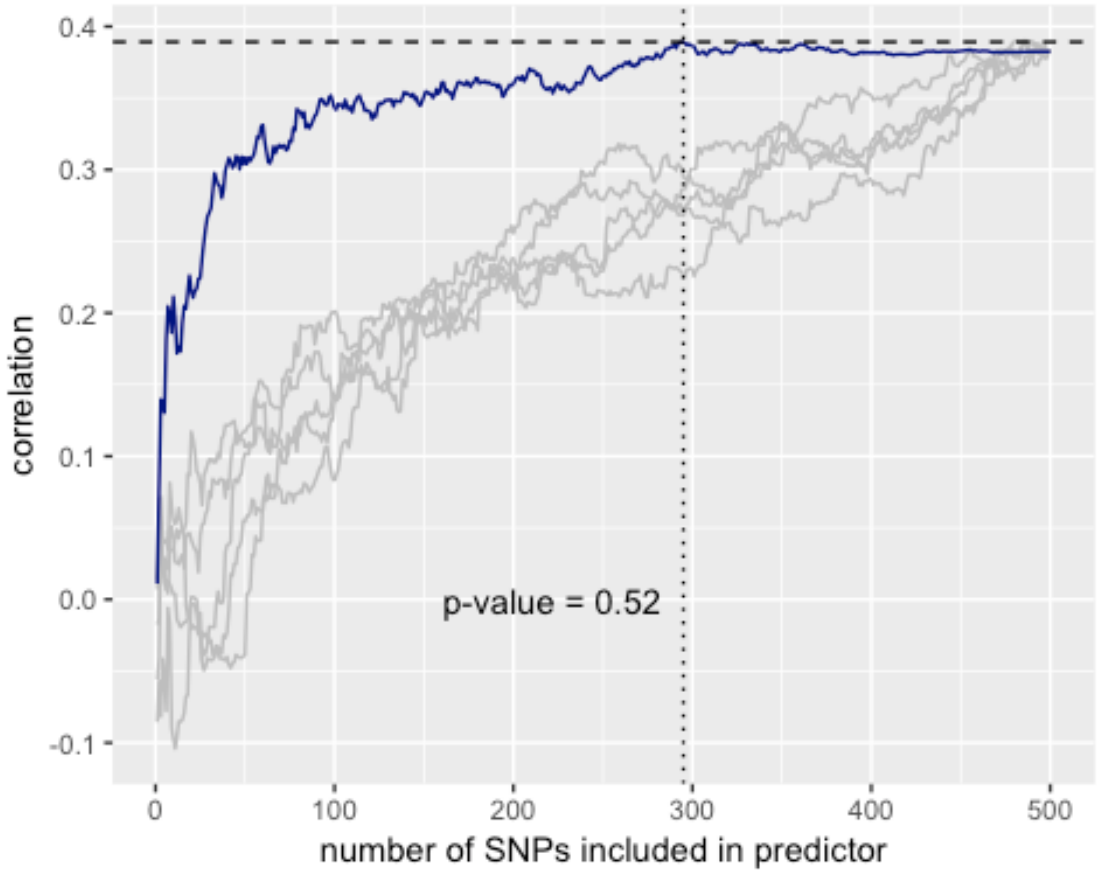

Figure 69: Selecting SNPs based on p-value

The blue line represent predictors based on an increasing number of SNPs, where SNPs are ordered from lowest to highest p-value from left to right. In comparison the grey lines represent predictors based on increasing numbers of SNPs when the SNPs are randomly selected.

As the plot above illustrates, prediction accuracy randomly fluctuates with inclusion of more or fewer SNPs, so selecting a predictor based on the best p-value threshold can lead to inflated estimates of accuracy.

Bias-variance tradeoff

You would think the words "bias" and "variance" are already overloaded enough, but in this section they will get yet another set of meanings. From Wikipedia:

In statistics and machine learning, the bias-variance tradeoff (or dilemma) is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:

- The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting: modeling the random noise in the training data, rather than the intended outputs.

So that's another perspective on why BLUP predictors are more accurate than GWAS or OLS predictors: BLUP finds the right balance between the simple GWAS model which only considers one SNP at a time, and the complex multiple regression OLS model, which considers all SNPs at the same time without any safeguard against overfitting the data. The heritability quantifies how much error is in the model and therefore how close BLUP can go towards fully conditional OLS effects, without overfitting the data.

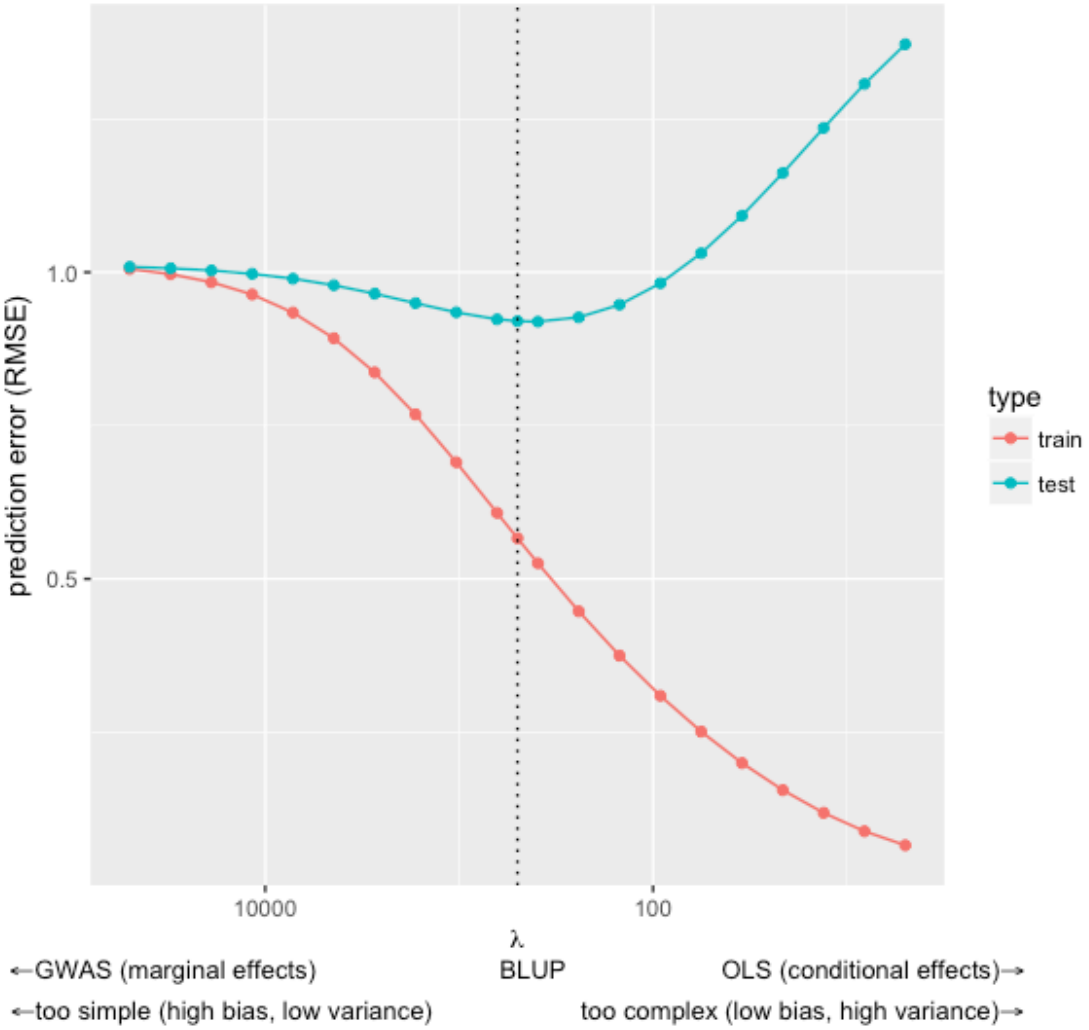

Figure 70: Model complexity of GWAS, BLUP and multiple regression OLS estimates

The lowest prediction error in the test set is achieved when the model is neither too simple nor too complex. Note that even at equal values of root mean square error (RMSE), predictions into the training set have much higher accuracy than predictions into the test set.

Estimation of variance components

Our model assumes that the phenotypic variance is composed of a genetic component and the rest (environment + error).

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2$$

In practice, we can only measure σ_y^2, and have to estimate the other components. This is equivalent to estimating the heritability (SNP-heritability, but here we don't make this distinction), since

$$h^2 = \frac{\sigma_g^2}{\sigma_y^2}$$

If we had an infinite sample size, then $\hat{\beta}_{j,OLS}$ would be the same as β_j and we could simply estimate σ_g^2 as the sum of the variances explained by each SNP:

$$\sigma_g^2 = \sum^j \beta_j^{*2}\, var(\mathbf{X}_j^*) = \sum^j \beta_j^2$$

```
c(h2, sum(beta^2 * varx))
## [1] 0.5000000 0.5184451
```

We can't do the same with finite sample size OLS estimates of β. Even though they are unbiased in the sense that $\mathbb{E}[\hat{\beta}_{OLS} \mid \beta] = \beta$, when we take the square of $\hat{\beta}_{OLS}$, it will be inflated by the estimation error, and so the σ_g^2 or h^2 estimate will be too large:

```
c(h2, sum(beta_ols^2 * varx))
## [1] 0.50000 3.12752
```

What if we had infinite sample size GWAS estimates of β? We still couldn't estimate σ_g^2 as $\sum^j \hat{\beta}^2_{j,GWAS}$, because GWAS estimates are marginal effect estimates, so if two SNPs are in high LD, their effect would be counted twice, whereas OLS estimates are conditional on all other SNPs, so their effect wouldn't be counted twice. In other words, $\hat{\beta}_{GWAS}$ is not an unbiased estimator of β (see section about variance of GWAS effects).

Another way to look at h^2 is as the variance in y that is explained by g. For any two variables, the variance explained in one by the other is given by R^2, or the squared correlation coefficient. For g and y this is defined as

$$h^2 = R^2_{\mathbf{g},\mathbf{y}} = \frac{cov^2(\mathbf{g},\mathbf{y})}{var(\mathbf{g})var(\mathbf{y})}$$

$cov(\mathbf{g},\mathbf{y})$ is the same as $cov(\mathbf{g},\mathbf{g}+\mathbf{e})$, and this is the same as $cov(\mathbf{g},\mathbf{g}) + cov(\mathbf{g},\mathbf{e})$. Further, $cov(\mathbf{g},\mathbf{g})$ is the same as $var(\mathbf{g})$, and our model assumes that $\mathbf{g}$ and $\mathbf{e}$ are uncorrelated, so $cov(\mathbf{g},\mathbf{e}) = 0$. Therefore,

$$cov(\mathbf{g},\mathbf{y}) = var(\mathbf{g})$$

and the equation above simplifies to

$$h^2 = R^2_{\mathbf{g},\mathbf{y}} = \frac{var(\mathbf{g})}{var(\mathbf{y})} = \frac{cov(\mathbf{g},\mathbf{y})}{var(\mathbf{y})}$$

Since the regression slope of y on x is defined as $\frac{cov(x,y)}{var(x)}$, the heritability is the same as the slope in a regression of $\mathbf{g}$ on $\mathbf{y}$. This is true by definition, but doesn't help us in estimating h^2 because we don't know $cov(\mathbf{g},\mathbf{y})$ any more than we know $var(\mathbf{g})$.

However, $var(\mathbf{g})$ can be estimated by looking at close relatives. The first attempt at estimating heritability was a regression of height measurements for a number of individuals on the average height of the parents of those individuals. A problem with this method is that it cannot distinguish between genetic and shared environmental effects. There are many other methods to estimate heritability that are based on close relatives, but here we will focus on methods that utilize unrelated individuals.

Variance explained per SNP

The genetic variance can be further partitioned into the variances explained by each SNP:

$$var(\mathbf{y}) = var(\mathbf{g}_j + \mathbf{g}_{rest} + \mathbf{e}) = var(\mathbf{X}_j\beta_j + \mathbf{g}_{rest} + \mathbf{e})$$

If we assume that not only g and e are uncorrelated, but also g_{j1} and g_{j2} (which we can, because we assume that all β_j are i.i.d.), then the variance explained by a particular SNP is

$$cor^2(\mathbf{g}_j, \mathbf{y}) = var(\mathbf{g}_j) = var(\mathbf{X}_j\beta_j) = \beta_j^2 var(\mathbf{X}_j) = \beta_j^2 = \beta_j^{*2} \times 2p_j(1-p_j)$$

The next parts present different methods to estimate h^2. Haseman-Elston regression and GREML require a GRM and phenotype data, whereas LD score regression requires GWAS effect estimates and LD scores.

Haseman-Elston regression

One simple way to estimate h^2 is to regress pairwise similarities of phenotypes on pairwise similarities of genotypes. For N individuals, there are $N \times (N-1)$ such pairs.

The genetic similarity for two individuals $i1$ and $i2$ is usually calculated as $\frac{\mathbf{X}_{i1}\mathbf{X}_{i2}^{\top}}{M}$, which are elements of the GRM.

There are two versions of HE-regression, with two different ways to calculate the phenotypic similarity. The first is to take the (negative) squared difference between phenotypes in the two individuals $(y_{i1} - y_{i2})^2$, the other is to take the product of the phenotypes $(y_{i1} \times y_{i2})$.

```
he_regression_v1 = function(A, y) {
  # estimates h2
  # A: GRM
  # y: phenotypes
  pairs = t(combn(1:length(y), 2))
  phenosim = y[pairs[,1]] * y[pairs[,2]]
  genosim = A[lower.tri(A)]
  unname(lm(phenosim ~ genosim)$coefficients[2])
}
```

```r
he_regression_v2 = function(A, y) {
  # estimates h2
  # A: GRM
  # y: phenotypes
  pairs = t(combn(1:length(y), 2))
  phenosim = -(y[pairs[,1]] - y[pairs[,2]])^2
  genosim = A[lower.tri(A)]
  reg = lm(phenosim ~ genosim)$coefficients
  unname(-reg[2]/reg[1])
}


h2_he_prod = he_regression_v1(grm, y)
h2_he_diff = he_regression_v2(grm, y)

c(h2, h2_he_prod, h2_he_diff)
## [1] 0.5000000 0.5309630 0.5256482
```

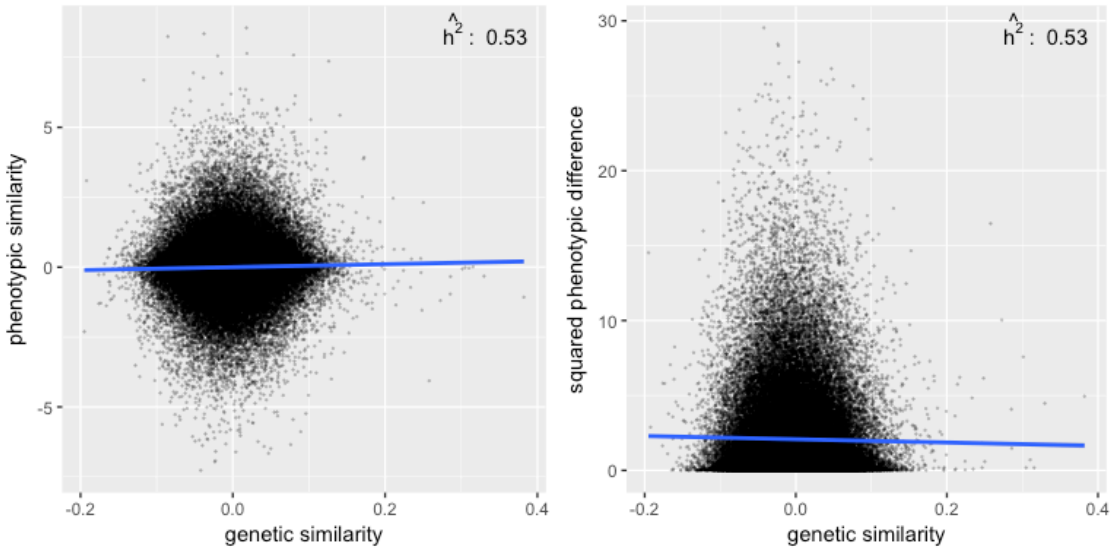

Figure 71: Two version of Haseman-Elston regression

GREML

HE regression is quick and simple, but suffers from large standard errors, so the estimates will often not be very accurate. A better way to estimate variance components is through GRM residual maximum likelihood estimation (GREML).

GREML directly builds on the phenotype model specified earlier:

$$Var[\mathbf{y}] = \mathbf{V} = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}_N$$

It runs an optimization algorithm finds the values of σ_g^2 and σ_e^2 that maximize the likelihood of this model. The likelihood of a model with specific parameters (σ_g^2 and σ_e^2) is defined as the probability of the data ($\mathbf{A}$ and $\mathbf{y}$) under this model :

$$\mathcal{L}(\sigma_g^2, \sigma_e^2 \mid \mathbf{A}, \mathbf{y}) = P(\mathbf{A}, \mathbf{y} \mid \sigma_g^2, \sigma_e^2)$$

Maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood, because the logarithm is a monotonically increasing function. The log likelihood of our model is:

$$log\mathcal{L}(\sigma_g^2, \sigma_e^2 \mid \mathbf{A}, \mathbf{y}) = -\frac{1}{2}(const + log|\mathbf{V}| + log|\mathbf{1}\mathbf{V}^{-1}\mathbf{1}^\top| + \mathbf{y}^\top\mathbf{P}\mathbf{y})$$

where $|\mathbf{V}|$ is the determinant of $\mathbf{V}$ and $\mathbf{P}$ is a projection matrix defined as

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{1}(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^\top\mathbf{V}^{-1}$$

Bear in mind that the description here is simplified a lot because we assume that there are no fixed effects.

Here, our model only consists of the two variance components σ_g^2 and σ_e^2, so the likelihood can be visualized as a heat map:

```
loglikelihood = function(sigma_g, sigma_e, grm, y, reml=TRUE) {
  V = sigma_g*grm + sigma_e*diag(nrow(grm))
  Vi = solve(V)
  if(reml) {
    P = Vi - (colSums(Vi) %*% t(colSums(Vi))) / sum(Vi)
  } else {
```

```
    P = Vi
  }
  -0.5*(log(det(V)) + log(sum(Vi)) + t(y) %*% P %*% y)
}
```

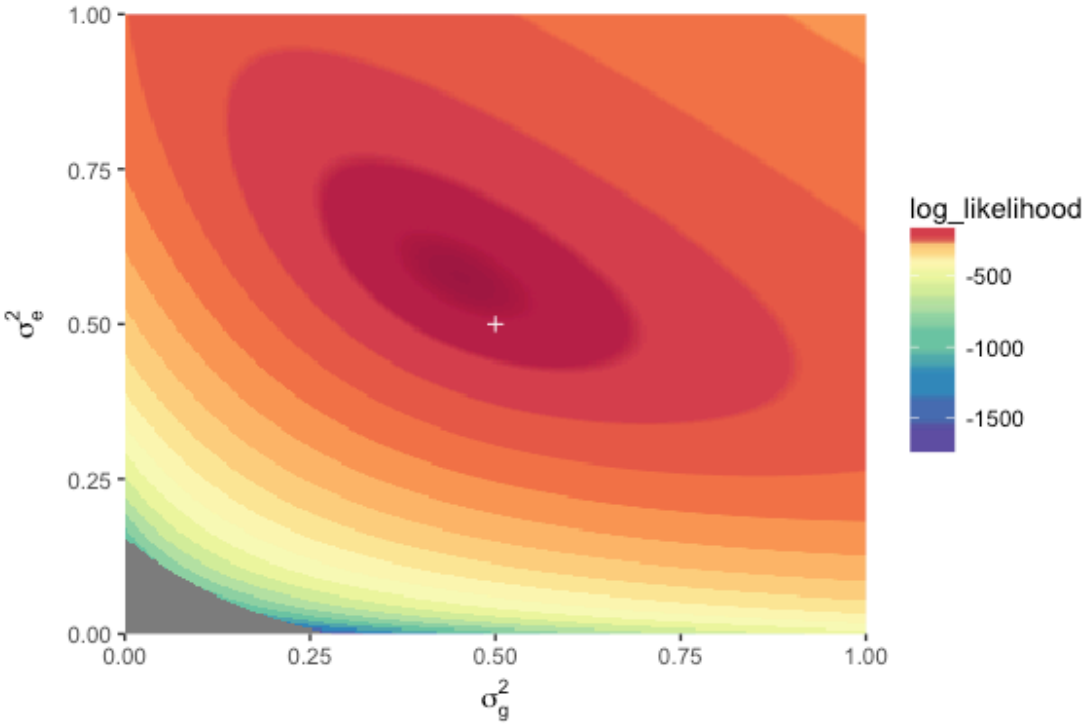

Figure 72: The log likelihood function as a heat map

To find the values of σ_g^2 and σ_e^2 that maximize the log likelihood function, we can use the Newton optimization method. This method requires first order and second order derivatives of the likelihood function to iteratively calculate δ values, so that the new estimates $\sigma_g^2 + \delta_g$ and $\sigma_e^2 + \delta_e$ have a higher likelihood than the present estimates σ_g^2 and σ_e^2.

The computation of the second order derivative matrix (the Hessian matrix) is expensive, but it can be approximated by the average information matrix, which is why the procedure below is called the average information (AI) algorithm.

The difference between REML and normal maximum likelihood estimation lies in the definition of the likelihood function. When using the normal maximum likelihood function, the inclusion of fixed effects will lead to biased estimates of the variance components. Since we don't model any fixed effects here, we wouldn't see a big difference here between ML and REML.

Below is a very simple example of this method based on stripped down GCTA code, which uses the average information algorithm for optimization. It iteratively updates estimates of σ_g^2 and σ_e^2 until it converges (or until this case, for a fixed number of iterations).

```
greml = function(A, y, varcmp=c(.95, .05), itermax=25) {
  # estimates g and e variance components
  # A: GRM
  # y: phenotypes
  out = varcmp
  for(iter in 1:itermax) {
    V = varcmp[1]*A + varcmp[2]*diag(nrow(A))
    Vi = solve(V)
    P = Vi - (colSums(Vi) %*% t(colSums(Vi))) / sum(Vi)

    Py = P %*% y
    APy = A %*% Py
    PPy = P %*% Py

    H = diag(2)
    H[1,1] = t(APy) %*% P %*% APy
    H[2,2] = t(Py) %*% PPy
    H[1,2] = H[2,1] = t(APy) %*% PPy
    Hi = solve(H + diag(2)*1e-6)

    R1 = t(Py) %*% APy - sum(P * A)
    R2 = t(Py) %*% Py - sum(diag(P))

    delta = Hi %*% c(R1, R2)
    varcmp = varcmp + 0.316*delta
    varcmp = pmax(1e-3, varcmp) # avoid negative var component
    out = rbind(out, varcmp)

    cat(paste0('Iteration ', iter, ': sigma_g = ', round(varcmp[1], 3),
               ', sigma_e = ', round(varcmp[2], 3),
               ', h2 = ', round(varcmp[1]/sum(varcmp), 3), '\n' ))
  }
```

```
  colnames(out) = c('sigma_g', 'sigma_e')
  out
}

varcmp = greml(grm, y)
## Iteration 1: sigma_g = 0.982, sigma_e = 0.074, h2 = 0.93
## Iteration 2: sigma_g = 0.975, sigma_e = 0.106, h2 = 0.902
## Iteration 3: sigma_g = 0.933, sigma_e = 0.148, h2 = 0.863
## Iteration 4: sigma_g = 0.866, sigma_e = 0.198, h2 = 0.814
## Iteration 5: sigma_g = 0.788, sigma_e = 0.255, h2 = 0.756
## Iteration 6: sigma_g = 0.713, sigma_e = 0.313, h2 = 0.695
## Iteration 7: sigma_g = 0.647, sigma_e = 0.369, h2 = 0.637
## Iteration 8: sigma_g = 0.594, sigma_e = 0.419, h2 = 0.586
## Iteration 9: sigma_g = 0.553, sigma_e = 0.46, h2 = 0.546
## Iteration 10: sigma_g = 0.523, sigma_e = 0.492, h2 = 0.515
## Iteration 11: sigma_g = 0.501, sigma_e = 0.517, h2 = 0.493
## Iteration 12: sigma_g = 0.486, sigma_e = 0.534, h2 = 0.476
## Iteration 13: sigma_g = 0.476, sigma_e = 0.547, h2 = 0.465
## Iteration 14: sigma_g = 0.469, sigma_e = 0.556, h2 = 0.457
## Iteration 15: sigma_g = 0.464, sigma_e = 0.562, h2 = 0.452
## Iteration 16: sigma_g = 0.461, sigma_e = 0.566, h2 = 0.449
## Iteration 17: sigma_g = 0.459, sigma_e = 0.569, h2 = 0.446
## Iteration 18: sigma_g = 0.457, sigma_e = 0.571, h2 = 0.444
## Iteration 19: sigma_g = 0.456, sigma_e = 0.573, h2 = 0.443
## Iteration 20: sigma_g = 0.455, sigma_e = 0.574, h2 = 0.443
## Iteration 21: sigma_g = 0.455, sigma_e = 0.574, h2 = 0.442
## Iteration 22: sigma_g = 0.455, sigma_e = 0.575, h2 = 0.442
## Iteration 23: sigma_g = 0.455, sigma_e = 0.575, h2 = 0.442
## Iteration 24: sigma_g = 0.454, sigma_e = 0.575, h2 = 0.441
## Iteration 25: sigma_g = 0.454, sigma_e = 0.575, h2 = 0.441
dat = data.frame(varcmp, log_likelihood=mean(ll[,3],na.rm=TRUE),
                 step=1:nrow(varcmp), row.names=NULL)
val = round(varcmp[nrow(varcmp), 1]/sum(varcmp[nrow(varcmp),]), 2)
heat + geom_point(data=dat[1:(nrow(dat)-10),], size=.5) +
  geom_path(data=dat,
            aes(sigma_g, sigma_e), colour='grey20', size=.1,
            arrow = arrow(angle = 25, ends = "last", type = "closed",
```

```
length=unit(5, 'mm'))) +
  annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
           label=as.character(paste0(bquote(hat(h^2)~": "~ ~.(val)),
collapse=''')), parse=T)
```

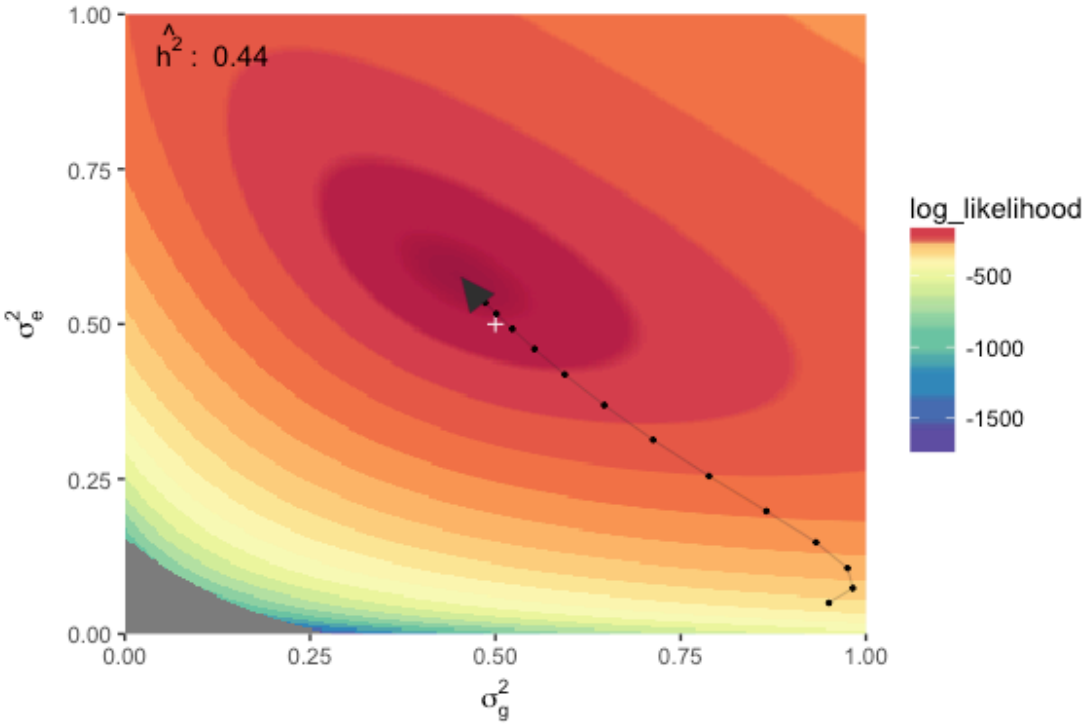

Figure 73: GREML finds the maximum in the log likelihood function

LD score regression

We have previously stated that the variance of $\hat{\beta}_{j,GWAS}$ can be approximated as

$$var(\hat{\beta}_{j,GWAS}) = l_j \frac{h^2}{M} + \frac{1}{N}$$

Because $var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ and $\mathbb{E}[\hat{\beta}_{j,GWAS}] = 0$, $var(\hat{\beta}_{j,GWAS}) = \mathbb{E}[\hat{\beta}^2_{j,GWAS}]$ and so

$$\mathbb{E}[\hat{\beta}^2_{j,GWAS}] = l_j \frac{h^2}{M} + \frac{1}{N}$$

Since $\hat{\chi}^2_j = \hat{z}^2_j = N\,\hat{\beta}^2_{j,GWAS}$, it follows that

$$\mathbb{E}[\chi_j^2] = \frac{Nh^2}{M} l_j + 1$$

Intuitively, this is because GWAS estimates are marginal effect estimates and this means that SNPs in high LD will have similar effect estimates, even if their true effect sizes are very different. A single large effect SNP within a group of correlated SNPs can give each of the correlated SNPs a large effect estimate. If a SNP is in high LD with many other SNPs, its chances of picking up the signal of another SNP are increased. This means that a SNP's LD score should be correlated to its GWAS effect estimate, and the correlation should be proportional to the heritability of the trait in question.

From this it follows that h^2 can be estimated as the slope in a regression of $\hat{\chi}_j^2$ on l_j:

$$\hat{h^2} = \frac{cov(\hat{\boldsymbol{\chi}}^2, \mathbf{l})}{var(\mathbf{l})} \frac{M}{N}$$

What is not shown here, is that the intercept in this regression is informative as well, because it will be higher when the effect estimates are inflated due to population stratification rather than real effects.

The regression estimate can be made more efficient by constraining the intercept (if we assume there is no population stratification) and by weighting the regression by the inverse of the LD scores.

```
ldsc = function(ldscores, beta01_gwas, n) {
  m = length(ldscores)
  chi2 = beta01_gwas^2 * n
  summary(lm(chi2-1 ~ ldscores + 0, weights=1/ldscores))$coefficients[1,1] *
m/n
}

ldsc(ldscores, beta01_gwas, n)
## [1] 0.5896047
ldsc(ldscoresld, beta01_gwasld, n)
## [1] 0.3335093
```

This probably didn't give a great estimates of heritability, because this is based on a very small sample size, and in the first case, almost no variance among the LD scores for different SNPs. Let's try it again with a larger data set with LD:

```
mm = 20000
nn = 5000
mmunique = sqrt(mm*2)
maf_block = runif(mmunique, 0, .5)
x012_block = t(replicate(nn, rbinom(2*mmunique, 2, c(maf_block, maf_block))))
polymorphic = apply(x012_block, 2, var) > 0
x012_block = x012_block[,polymorphic][,1:mmunique]
maf_block = c(maf_block, maf_block)[polymorphic][1:mmunique]

# create haplotype blocks, so there is some variance in ld-scores
x012_block = jitter(x012_block[,rep(1:mmunique,1:mmunique)[1:mm]], .4)
x01_block = scale(x012_block)

# we can approximate ld-scores from how we simulated LD
ldscores_block = rep(1:mmunique,1:mmunique)[1:mm]

m_block = ncol(x01_block)
beta01_block = rnorm(m_block, 0, sqrt(h2/m_block))
g_block = x01_block %*% beta01_block
y_block = g_block + rnorm(nn, 0, sqrt(1-h2))

beta01_gwas_block = t(x01_block) %*% y_block / nn

h2_est = ldsc(ldscores_block, beta01_gwas_block, nn)
h2_est
## [1] 0.456884
```

This should be closer to the true value of h^2. While this may not be as precise as the estimates from the other methods, LD score regression has the advantage of requiring only summary statistics data (effect estimates), not individual level genotype data. The LD scores can be estimated in a reference population.

```
chi2 = beta01_gwas_block^2 * nn
sl = summary(lm(chi2-1 ~ ldscores_block + 0,
weights=1/ldscores_block))$coefficients[1,1]

dat = data.frame(x=ldscores_block, y=chi2,
                 gr=cut(ldscores_block, seq(min(ldscores_block),
max(ldscores_block), length=30)))
dat_groupmean = group_by(dat, gr) %>% summarise(x=mean(x), y=mean(y))

val = round(h2_est, 2)
layers = list(geom_abline(intercept=1, slope=sl, col='blue', size=1.0),
              xlab('LD scores'), ylab(expression(Chi^2)),
              scale_colour_continuous(low='black', high='green'),
              theme(legend.position = 'none'),
              annotate('text', -Inf, Inf, hjust=-.2, vjust=1.2,
                       label=as.character(paste0(bquote(hat(h^2)~": "~
~.(val)), collapse=''')), parse=T))
p1 = ggplot(dat, aes(x, y, col=x)) + geom_point(size=.01) + layers
p2 = ggplot(dat_groupmean, aes(x, y, col=x)) + geom_point() + layers

grid.arrange(p1, p2, ncol=2)
```

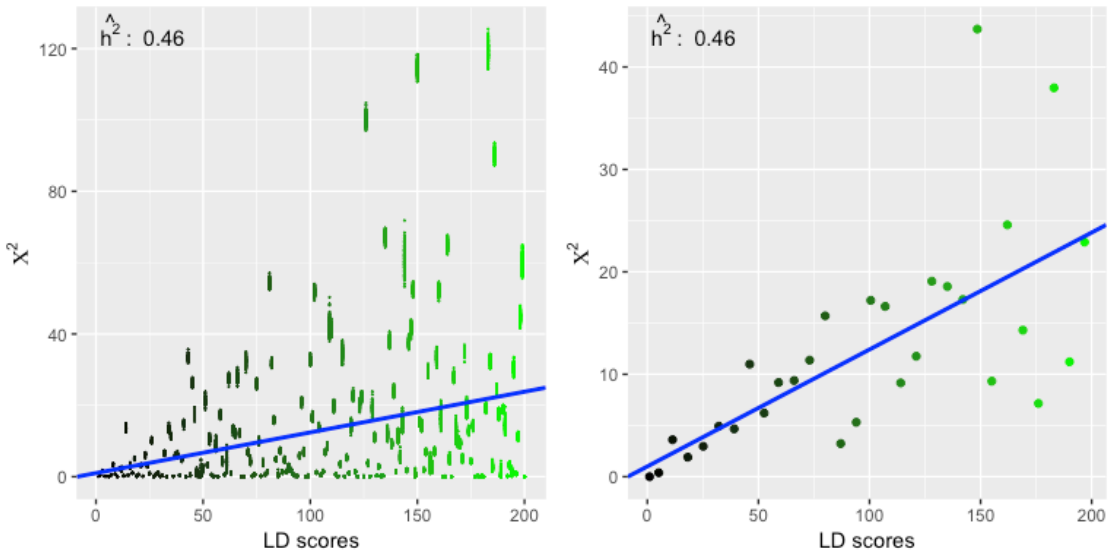

Figure 74: LD score regression visualized

The left panel shows the LD score regression for all SNPs. The right panel groups SNPs by LD score and shows the mean χ^2 value in each bin.