Polygenic methods and their application to psychiatric traits

Naomi R Wray[1*], S Hong Lee[1], Divya Mehta[1], Anna AE Vinkhuyzen[1], Frank Dudbridge[2], Christel M Middeldorp[3,4]

1. The University of Queensland, Queensland Brain Institute, Queensland, Australia 4068
2. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
3. Department of Biological Psychology, Neuroscience Campus Amsterdam, VU University Amsterdam, The Netherlands
4. Department of Child and Adolescent Psychiatry, GGZ inGeest/VU University Medical Center, Amsterdam, The Netherlands

* Corresponding author. Naomi.wray@uq.edu.au

Abstract

Background Despite evidence from twin and family studies for an important contribution of genetic factors to both childhood and adult onset psychiatric disorders, identifying robustly associated specific DNA variants has proved challenging. In the pre-genomics era the genetic architecture (number, frequency and effect size of risk variants) of complex genetic disorders was unknown. Empirical evidence for the genetic architecture of psychiatric disorders is emerging from the genetic studies of the last five years.

Methods and scope We review the methods investigating the polygenic nature of complex disorders. We provide mini-guides to genomic profile (or polygenic) risk scoring and to estimation of variance (or heritability) from common SNPs. We review results of applications of the methods to psychiatric disorders and related traits and consider how these methods inform on missing heritability, hidden heritability and still-missing heritability.

Results Genome-wide genotyping and sequencing studies are providing evidence that psychiatric disorders are truly polygenic, that is they have a genetic architecture of many genetic variants, including risk variants that are both common and rare in the population. Sample sizes published to date are mostly underpowered to detect effect sizes of the magnitude presented by nature, and these effect sizes may be constrained by the biological validity of the diagnostic constructs.

Conclusions Increasing the sample size for genome wide association studies of psychiatric disorders will lead to the identification of more associated genetic variants, as already found for schizophrenia. These loci provide the starting point of functional analyses that might eventually lead to new prevention and treatment options and to improved biological validity of diagnostic constructs. Polygenic analyses will contribute further to our understanding of complex genetic traits as sample sizes increase and as sample resources become richer in phenotypic descriptors, both in terms of clinical symptoms and of non-genetic risk factors.

Key points:

- Genome-wide association data provide evidence for the polygenic architecture of psychiatric disorders and traits, i.e. these traits are influenced by many genetic variants and affected individuals may carry a polygenic burden of risk alleles.
- We provide mini-guides for polygenic methods of genomic profile (or polygenic) risk scoring and of estimation of variance (or heritability) from common SNPs.
- Polygenic methods applied to currently available samples provide evidence that part of the missing heritability is just hidden and that with increasing sample sizes the number of genome-wide significant hits will substantially increase, as already achieved for schizophrenia.
- The identification of genetic variants provides the starting point of functional analyses that might eventually lead to new prevention and treatment options and to improved biological validity of diagnostic constructs.

Introduction

Twin and family studies have reported a significant contribution of genetic factors to both childhood and adulthood onset psychiatric symptoms and disorders, with heritability estimates in the range of 0.4-0.8 (**Figure 1**), implying that inherited DNA variants are important in the etiology of these disorders. Therefore, identification of specific genetic variants has been a research goal for some decades as a mechanism to gain insight into etiology. Recent advances in technology have allowed the systematic testing of genetic variants across the genome for association with traits measured on unrelated individuals. The aims and outcomes of these genome-wide association studies (GWAS) have been reviewed elsewhere[1,2]. Briefly, in GWAS the genetic variants tested are single nucleotide polymorphisms (SNPs) or large copy number variants (CNVs; submicroscopic insertions/deletions, usually > 100kb); this review considers analyses from the SNPs. Each SNP is tested for association with the trait, which is difference in mean score of a quantitative trait for the alternate SNP alleles, or differences in allele frequencies between cases and controls in the analysis of disease traits. The SNPs measured are common genetic variants with a minor allele frequency of at least 0.01 and mostly higher. About a 1 million independent association tests are conducted and hence, to avoid chance findings, the threshold for declaring significance of an association test is 0.05/1 million or $5x10^{-8}$ (ref[3]). The correlation structure of the genome means that each SNP tested is correlated with many other DNA variants within a ~ 1MB region (**linkage disequilibrium**). Thus, an associated SNP is unlikely itself to be the risk conferring variant but tags a risk region for follow-up study. Given the disappointing results of association studies conducted prior to the GWAS era in which hypothesis driven candidate genes were tested, there were high hopes that the hypothesis-free systematic evaluation of the whole genome in GWAS would enable identification of genetic variants associated with psychiatric disorders (and other **complex genetic traits**). The first empirical data came from the GWAS of the Wellcome Trust Case Control Consortium (WTCCC)[4] that benchmarks the beginning of the GWAS era. Across all seven disorders studied (including bipolar disorder) each of ~2000 cases with 3000 shared controls, 14 independent loci surpassed the significance threshold, but these loci explained only a small proportion of heritability. The first phase of GWAS for the major psychiatric disorders schizophrenia, bipolar disorder, major depressive disorder (MDD), autism spectrum disorders (ASD) and attention deficit hyperactivity disorder (ADHD) [4-10] showed a similar picture with few or no genome wide significant hits, indicating that common variants of large effect are not part of nature's repertoire. All studies had excellent power to detect genetic variants with an odds ratio of 1.5 and a minor allele frequency of 0.2 and reasonable power to detect an effect with an odds ratio of 1.3. The absence of significant results raised the question whether common variants are of sufficient relevance in the development of psychiatric disorders to pursue with GWAS. Here we review methods that use GWAS to provide critical empirical evidence of an important **polygenic** contribution to common psychiatric disorders as was first proposed 45 years ago[11]. In this review, we first define heritability and related measures. We then consider the methods that have demonstrated the evidence of a polygenic contribution to the genetic architecture of complex traits, diseases and disorders including psychiatric disorders. Next, we review applications of these methods to psychiatric disorders and related phenotypes. Lastly, we draw conclusions and implications for future research. In this edition Thapar and Gordon contribute a Perspectives article providing further clinical interpretation of these methods[12].

Heritability

Evidence for a genetic contribution to psychiatric disorders comes from the consistently reported 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 the increased 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. 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 than comparison of risks to relatives. **Figure 1** shows heritability for a range of psychiatric disorders. 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 on the liability scale depend on knowledge of baseline risk of disease in the population from which the twin and family cohorts are drawn, and estimates of baseline risk are often surprisingly difficult to pin down. They may also vary between populations, across ages and may depend on whether non-genetic factors have been recorded and included in the analysis. Hence, in reality heritability estimates should be viewed as pragmatic benchmarks representing evidence for low, moderate or high contributions of genetic effects.

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). 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.

Missing heritability in GWAS

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 genetic variants to variance can be calculated on the liability scale[13-15] to allow direct comparison of the contribution to risk of each locus on the same scale as heritability is reported. Assuming independence, 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. This is denoted by 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 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. As mentioned above, these GWAS yielded few GWS results with h^2_{GWS} much less than h^2. This difference has been termed “missing heritability”[16]. As sample sizes have increased, the number of GWS variants have increased for both quantitative traits and diseases (see Fig 2 in Visscher et al[2]) providing empirical evidence that common variants do play a role in **complex genetic traits**. Currently, GWS variants explain < 0.02 of variance for bipolar disorder, MDD, ADHD and ASD and ~0.07 for schizophrenia. An exception is Alzheimer’s disease in which ~0.18 of variance is explained, but this is mostly attributable to variants of the APOE gene, identified in the pre-GWAS era.

The observed increase in number of significant results for the traits for which larger sample sizes have been accumulated, 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 traits? We describe two methods that were developed to investigate the polygenic architecture of traits using data sets that are currently available. Although these data sets may be underpowered to detect the individual small effects as GWS they can provide evidence for contributions of common variants of small effect sizes as explained in the sections below by also taking into account contributions from SNPs that do not reach genome-wide significance. These analyses increase our understanding of the genetic architecture of traits to which they are applied and provide empirical evidence to help decisions about future experimental design.

Genomic profile risk scoring

Many GWAS that report no or few GWS associated SNPs show more small p-values of association than expected by chance. This tell-tale sign of a polygenic genetic architecture provided stimulus for the development of methods that capture this signature. The first such method was profile scoring (see **Box 1**). Briefly, in its standard application, a GWAS is conducted in a sample denoted the "discovery" sample. The risk alleles and their effect sizes are then used to generate genomic profile risk scores (GPRSs) in an independent "target" sample, using SNPs whose p-values in the discovery sample are below some threshold (**Box 1** step 5). A GPRS is calculated for each individual in the target sample as the sum of the count of risk alleles weighted by the effect size (log odds ratio for case-control) in the discovery sample. The profile score is evaluated through regression of the target phenotype on the GPRS after accounting for other known covariates. The target phenotype could differ from the phenotype in the discovery sample, allowing cross-phenotype analyses as we discuss later. In association analysis, the aim is to identify specific associated variants and the stringent threshold for declaring significance of individual SNPs is important, providing confidence that the identified variants are true positives. This is important because specific (and costly) follow-up studies are directed at loci that surpass this cut-off and hence false positives cannot be tolerated. In contrast, GPRS analyses aim to provide insight in the genetic architecture using evidence for association from variants that do not pass the stringent threshold of association. As the threshold of discovery sample p-value increases, the number of SNPs included in the GPRS increases and hence the ratio of false: true positives increases. However, profile score analyses can tolerate inclusion of some false positives since, on balance, useful information from the true positives may still contribute. Selection of SNPs into profile scores is therefore based on much less stringent p-value thresholds than in association analysis of single variants, and in principle all SNPs could be included into the score (see **Box 1** sample size considerations).

For quantitative traits, the variance of the phenotype explained by the GPRS is expressed by the regression $R^{2,}$ that is, the squared correlation between the trait and the GPRS. Given the size of target samples, a small R^2 can be highly significant. For disease traits, following Purcell et al[8], the scaled pseudo-R^2 from logistic regression, Nagelkerke's R^2 (NR^2) is often reported. However, the NR^2 is difficult to benchmark since it depends on the proportion of cases in the target sample, and is not on the liability scale, so cannot be directly compared to heritability estimated from twin or family data. Alternative statistics of efficacy of risk prediction are described and compared by Lee et al[17] and include variance explained on the liability scale, which can be directly compared to h^2.

Factors that can bias GPRS results have been discussed in detail elsewhere[18]. An important consideration is the delineation of the discovery and target samples. They should be independently collected and exclude close relatives. Another issue is the choice of SNPs used in the GPRS. SNPs located within the same genomic region are more likely to be inherited together i.e., the alleles are correlated and the SNPs are in **linkage disequilibrium** (LD). In practice, the most common strategy, following the initial publication[8], is to prune SNPs based on a p-value informed clumping algorithm and to choose a relatively stringent LD threshold (say $r^2 < 0.2$ across 500 kb). **Clumping** aims to select SNPs so that the most associated SNP in the region is selected into the SNP profile set. However, this is not an optimal strategy as the LD threshold selected is somewhat arbitrary so that multiple SNPs may be retained that show association generated by the same causal variant and at the same time correlated SNPs with associations driven by independent causal variants may be excluded.

We have described the GPRS method as it is most frequently applied in the psychiatric research literature. The method is simple, robust and intuitive. However, other strategies for creation of individual risks may be more optimal (e.g.,[19,20]). A full discussion of these methods is beyond the scope of this review.

Estimating variance explained by all SNPs

In standard association analysis the effect of SNPs are tested one at a time. Since the number of genome-wide SNPs is currently greater than the number of individuals in the study and since correlated SNPs will show correlated association results, simple addition of variance explained by each SNP may overestimate the true total variance explained by all SNPs. This can be overcome by LD pruning of the SNP set, but the choice of LD threshold will be arbitrary and can influence the results. Therefore, other methods have been developed that analyse all SNPs simultaneously, first for quantitative traits[21-25], and later extended to disease traits[26]**.** These methods have been called GREML[27] (for genomic-relationship-matrix restricted maximum likelihood) and are implemented in the software GCTA[28], see **Box 2**. Briefly, the method uses the genome-wide markers to estimate the genetic similarity between individuals in the study who are **conventionally unrelated**. The variance explained by all SNPs (h^2_{SNP}) is estimated to be greater than zero when genetically more similar individuals are phenotypically more similar, or in a case-control design when cases are genetically more similar to each other than they are to controls. Stringent QC[26] is needed to ensure that artefacts do not bias the estimates, a problem much more likely in analysis of disease traits than quantitative traits. As part of the QC process closely related individuals are removed. These are detected from the genotype data as large genetic similarities between pairs of individuals. Removal of close relatives ensures that estimates reflect the tagging of causal variants through population **LD**. If more closely related individuals were to be included the estimate would reflect the much higher LD in family members. Moreover, by using only individuals **conventionally unrelated** the estimates of h^2_{SNP} are unlikely to be contaminated by common environmental effects that can bias estimates of heritability from family data.

In GREML, contributions from any specific locus are not evaluated. A significant h^2_{SNP} when few genome-wide significant associated SNPs have been identified provides direct empirical support from currently available data that increasing sample size is a worthy research objective. We emphasise that h^2_{SNP} is not expected to be as large as h^2, since h^2_{SNP} only reflects variants correlated with the common SNPs included on genome-wide SNP chips. This is because it is not possible for rare and uncommon genetic variants to be highly correlated with common SNPs[29] and so their contribution to $h^2_{SNP}$is limited. In contrast, rare and uncommon variants are shared between family members and so contribute to estimates of h^2 recognising that different families may have different rare variants segregating. The difference between h^2_{SNP} and h^2 can provide insight into genetic architecture in terms of the relative importance of common variants, which may differ between traits (**Figure 1**).

A bivariate model[30] allows estimation of the SNP-heritability of two traits and the SNP-correlation ($r_{g\text{-}SNP}$) between them. Some thought is needed for the interpretation of the SNP-correlation. Firstly, the correlation reflects the average genome-wide relationship between two disorders. For example, a zero SNP-correlation could result either from no relationship at all between the disorders, or from positive correlations at some genomic locations cancelled out by negative correlations at other locations. Calculating correlations for different functional categories of SNPs could identify genomic locations that show different directions of sharing between disorders. Secondly, the extent to which the SNP-correlation reflects the genetic correlation estimated from family data depends on the unknown underlying genetic architecture as to whether the SNP-correlation estimated from common SNPs is the same as the correlation across the whole allelic spectrum. The SNP-coheritability (i.e. $r_{g\text{-}SNP}h_{SNP\text{-}1}h_{SNP\text{-}2}$) allows direct comparison of the relationship between disorders on the same scale as the SNP-heritabilities. Bivariate methods can be applied to data sets in which all individuals are measured for both traits, but in this case inflation of estimates of genetic sharing by sharing of environmental risk factors may be difficult to avoid. More interesting is the application of bivariate methods to independent data sets each measured for a different trait, so that sharing between individuals reflects only genetic factors[30].

While the methods that estimate h^2_{SNP} from genome-wide SNP data considered jointly in a single analysis are statistically optimal, the analyses can be time consuming and computationally demanding due to the calculation of genomic relationship matrices. Moreover, the large sample sizes required often involve sharing of genotype data among research groups, which is not always possible (see sample size notes in **Box 2**). Therefore, approximate methods based on association summary statistics (of RAF, OR, p-value, sample size) are appealing. Under a polygenic model, test statistics of association are expected to be inflated compared to the distribution of test statistics in the absence of association, and h^2_{SNP} can be estimated directly from the mean test statistic[14,31-33] as well as from the results of GPRS association testing[32]. As discussed above, biases due to artefacts are a particular concern for case-control studies. When estimating h^2_{SNP} from the SNP data, QC strategies can investigate the potential of biases, but such strategies are not available for analyses based on summary statistics and so the potential for biases in these results should be recognised. The relationship between GPRS and GREML results are discussed in **Box 3**. Studies that apply multiple methods can help evaluate the validity of approximate methods.

Power and sample size

Prior to undertaking a polygenic analysis a power calculation establishes boundaries of what can be achieved. In GPRS, power depends on both the sample size of the discovery and target samples. Firstly, detecting a variance explained as being significantly different from zero depends on the sample size of the target sample. Secondly, the ability of the GPRS to explain variance in the target sample depends on the underlying genetic architecture of the disorder (unknown and not in our control) and on the sample size of the discovery sample to estimate accurately individual SNP effects. Once target sample sizes reach a reasonable size there is little to be gained in increasing them as they already have excellent power to detect a variance explained as different from zero. In contrast, increasing the discovery sample size will continue to increase the variance explained and the GPRS for each individual become more accurate, which is advantageous for other analyses (e.g. relating GPRS to sub-phenotypes). Only when researchers have access to genotype data for all samples can a choice be made about division into discovery and target samples. Dudbridge[34] provides a power calculator for GPRS and also a pragmatic rule of thumb under circumstances when the split into discovery and target can be chosen: discovery and target samples should be of equal size until the target sample is ~2000 cases and 2000 controls, and then the additional samples should be included into the discovery sample.

For GREML, power can be estimated from an online calculator[35] in which sample size (number of pairwise relationships) can directly predict the standard error of the estimate of h^2_{SNP} which is independent of the magnitude of the estimate. For a quantitative trait a sample size of 4500 is needed to have 80% power to detect h^2_{SNP} of 0.2 as being significantly different from zero. For case–control studies, power also depends on the proportion of cases in the sample and the risk of disease in the population. Samples of 4500 with equal proportion of cases and controls have at least 80% power to detect h^2_{SNP} of 0.2 as being significantly different from zero for disorders of disease risk 0.1 or less. In bivariate analyses the s.e. of the correlation depends on both the magnitude of the SNP-heritabilities for the two disorders and the magnitude of the correlation, as a rough rule of thumb samples of at least 5000 are needed for each of the two disorders for h^2_{SNP} of 0.2 and SNP-correlation of 0.2.

Polygenic analyses in psychiatry

There are three broad applications of polygenic analyses in psychiatry: single disorder analyses, cross disorder analyses and sub-phenotype analyses. Here we review the results of studies using any one of these applications. Studies were identified in Web of Knowledge (www.webofknowledge.com) by searching for studies that cited Purcell et al[8], Yang et al[21], Lee et al[26].

Single disorder analyses

Amongst the psychiatric disorders, schizophrenia is the flagship disorder achieving larger samples more quickly than the other disorders. In 2008, GWAS for schizophrenia[8,9] were published with sample sizes of ~3000 cases and identified only 1 genome-wide significant association (and also an excess of large rare copy number variants in cases[36,37]). At this point many considered GWAS in psychiatry a failure[38]. Polygenic analysis methods were central in demonstrating that the first phase of GWAS were underpowered, which propelled the drive for larger sample sizes that is now starting to pay off. We first consider polygenic results for schizophrenia and illustrate the relationship between the GPRS and GREML methods.

The first application of GPRS used the International Schizophrenia Consortium (ISC) data as the discovery sample and the Molecular Genetics of Schizophrenia (MGS) cohort as the target sample and gave a NR^2 of 0.032, which through simulation was shown to be consistent with h^2_{SNP} of 0.34, i.e.,34% of the variance in liability to schizophrenia is explained by many common SNPs of small effect. The approximation method[34] applied to these GPRS results gives h^2_{SNP} = 0.29. Application of GREML to the ISC data generated a direct estimate of h^2_{SNP} = 0.33 (reducing to h^2_{SNP} = 0.27 after stringent QC). These results are discussed in detail in **Box 3** and demonstrate some robustness in that different methods (although underpinned by similar theory) generate convergent results. The NR^2 of 0.032 is modest since many SNPs included in the GPRS do not contribute and only add noise. The simulations conducted by the ISC[8] (their Figure S8) suggested that as sample size increased there would be a better separation of true and false positives and an increase of NR^2. Application of GPRS using the Psychiatric Genomics Consortium data as discovery (8832 cases, 12067 controls) and the independent Swedish Schizophrenia sample as target (5001 cases, 6243 controls) generated NR^2=0.06, with maximum NR^2 achieved using SNPs with p-value threshold < 0.3, very much in line with predictions from the ISC simulations.

Table 1 and **Supplementary Table 2** give an overview of other studies that investigated the polygenic architecture of psychiatric traits and disorders using GPRS. The studies show a rather consistent pattern over the various phenotypes with significant predictions but low explained variance (between 0.001 and 0.03). The results, to date, are mostly less significant than in the schizophrenia studies, which reflects the more limited sample sizes available.

Univariate GREML analyses show that the variance explained by all SNPs (h^2_{SNP})(**Figure 1, Supplementary Table 1**) is mostly estimated at around 0.2 or higher for psychiatric disorders. Further insight into the genetic architecture is achieved by partitioning h^2_{SNP} based on SNP annotation such as based on chromosome, function and minor allele frequency[39]. Partitioning by chromosome confirmed the polygenic model with variance attributable to each chromosome being proportional to chromosome length. In contrast, for Alzheimer disease[40] significantly more variance was attributable to the chromosome 19, the genomic location of APOE. The variance explained by the subsets of SNPs based on minor allele frequency bin indicated that the variance attributable to SNPs must be explained, at least in part, by common causal variants (rather than common SNPs tagging only rare causal variants)[39]. Finally, SNPs in and around genes that are preferentially expressed in the brain explain a larger proportion of the variance than expected based on the proportion of the genome that they represent (0.3 of the variance explained versus 0.2 of the genome represented)[39,41]. Variance partitioning has been applied to schizophrenia[39], bipolar disorder[41], ASD[42,43], all generating qualitatively similar results. Application to Tourette Syndrome (TS) and OCD[44] using the same control set for each disorder provided evidence for a different genetic architecture between the disorders in which less common variants contributed more to h^2_{SNP} for TS than for OCD.

For psychiatric related quantitative traits the trend is towards smaller estimates of h^2_{SNP} compared to the psychiatric disorders. Analysis of a wide range of quantitative traits measured in ~2500 unrelated children from the Twins Early Development Study (TEDS) [45-47] showed significant h^2_{SNP} for height, weight and cognitive ability in line with those reported from other studies, but negligible h^2_{SNP} for childhood behavioural traits (anxiety, depression, hyperactivity, conduct) despite substantial estimates of heritability using the family data from which the samples were drawn. Similarly, estimates of h^2_{SNP} for neuroticism and extraversion from samples of ~12,000 unrelated individuals were 0.06 (s.e. 0.03) and 0.12 (s.e. 0.03) respectively[48]. These results may point to quantitative behavioral traits being composite genetic traits, such that family members score similarly (hence substantial estimates of heritability) but that different families may score similarly but for different genetically determined reasons (hence low estimates of h^2_{SNP}). More data sets are needed to explore this further, since other studies show higher estimates but with high standard errors, e.g., 0.26 (s.e. 0.12, meta-analysis of three sample estimates) for preschool internalising symptoms[49] and 0.18 (s.e. 0.07) for quantitative scores of social communication skills measured in a community sample[43]. The finding for social communication skills, related to ASD, is especially noteworthy given the evidence for the involvement of rare variants in ASD[50]. These results show that common variants are also of importance is ASD[42].

Bivariate analyses can be applied to two data sets of the same disorder, such analyses generate three estimates of h^2_{SNP}, one from each subset and one estimated between subsets (from the co-heritability). The estimate from the two samples combined into a single sample will be a weighted average of these three values. Such analyses explore the heterogeneity of GWAS data sets within a disorder, which can be summarised through the SNP-correlation. This correlation is expected to be 1 when the two data sets are of the same disorder and sampled from the same homogeneous population. SNP-correlations less than 1 could imply inflation of h^2_{SNP} from each data set relative to the h^2_{SNP} estimated between data sets reflecting genotyping artefacts or else data set specific variants. When PGC data sets were split into either 2-3 subsets, heterogeneity between estimates was much more evident for bipolar disorder, MDD, and ADHD than for schizophrenia and ASD (as shown by SNP-correlations **Figure 2**). Future studies designed to understand this observation that may reflect genetic and phenotypic heterogeneity implicit in diagnostic class may be critical to maximise power in GWAS (since the observed heterogeneity will also impact association analysis). Analyses have also been conducted for data of the same disorder from different ethnicities (**Figure 2**), for example the SNP-correlation between schizophrenia in European American ancestry vs African American ancestry was 0.63 (s.e. 0.22)[51] compared to the correlation of 0.83 (s.e. 0.09) between the European ISC and the European American MGS samples[51]. Likewise SNP-correlation between Chinese and European ADHD samples was estimated to be 0.39 (s.e. 0.15) compared to 0.71 (s.e. 0.17) between European ADHD sub-sets[52]. These analyses demonstrate that there are likely ancient common variants contributing to the etiology of these disorders, even though different LD structure and recent population specific causal variants generate lower correlations between ancestries than between sample subsets of the same ancestry. In general, these results provided strong support for a polygenic contribution to psychiatric disorders.

Cross-disorder analysis

The first application of cross-disorder polygenic analyses was with the International Schizophrenia Consortium as the discovery sample and the WTCCC bipolar disorder sample as the target[8], generating a NR^2 of 0.01 (ref[4]) with p-value of $1*10^{-12}$. Importantly the schizophrenia discovery sample did not significantly explain any variance in the other six non-psychiatric WTCCC traits, which all used the same set of controls. Genome-wide sharing between the schizophrenia and bipolar disorder cases is implicated. GPRS and bivariate GREML results have been reported between all five disorders in the Psychiatric Genomics Consortium study[41,53]. Applying the quantitative genetics theory[34] (**Box 3**) to the GPRS results provides estimates of the genetic correlation that agree well with the GREML results (**Figure 3**).

Studies investigating the overlap between disorders using GPRS are summarized in **Table 2** and **Supplementary Table 3**. On the whole, genetic relationships between disorders implied from GPRS and bivariate GREML agree in broad terms with expectations from twin and family studies, i.e., most studies show genetic overlap between disorders. However, some cross-disorder results are unexpected or inconsistent. For example, the genetic correlation of 0.43 between schizophrenia and MDD was surprising for many, but when translated to the expected increased risk to first-degree relatives of 1.6, this was found to be highly consistent with a meta-analysis of results from family studies (OR 1.5, 95% CI 1.2-1.8)[41]. That a genetic correlation of 0.43 translates into a modest increased risk to relatives may seem surprising but is a direct reflection that MDD is a common disorder. We discussed above that it is difficult to benchmark the genetic contribution to disease from risks to relatives. The lack of genetic overlap between ADHD and ASD (and also between these disorders and other disorders) from the GPRS and bivariate GREML analyses was unexpected since family studies point to a shared genetic background for ASD, ADHD and bipolar disorder[54-57]. As discussed above, sample sizes can impact on GPRS results and sample sizes for ADHD and ASD samples are small relative to other disorders. For example, GPRS generated from a schizophrenia discovery sample and applied to ASD target sample was not significant in the Vorstman et al[58], but did reach significance ($p<0.05$) when additional samples had accumulated[41]. In principle, GREML results should be unbiased regardless of sample size, with the standard error of the estimate decreasing with sample size. However, as discussed above, the estimates of h^2_{SNP} between and within subsets showed more heterogeneity between subsets for some disorders (**Figure 2**) than expected from the standard errors, which could imply phenotypic or genetic heterogeneity or artefacts and these could impact the GPRS and bivariate GREML results. Interestingly, GPRS analysis using the PGC schizophrenia and/or bipolar disorder sets as discovery samples and 727 ADHD cases and 2067 controls as target sample found the most significant association when SNPs with association p-value < 0.5 for both schizophrenia and bipolar disorder were used to generate the profile score SNP list[59]. These results point to important sharing of genetic risk factors between the disorders. However, the more significant result from the combined disorder discovery sample may also reflect increased power for disorder-shared variants through larger sample size. More data is needed to fully understand the relationship between disorders.

GPRS applied to personality traits to explore relationship with psychiatric disorders such as anxiety/depression and bipolar disorder suggest that neuroticism is related to anxiety/depression disorders or related traits (as expected from family studies), although results are somewhat inconsistent[60]. For extraversion, the picture is more complicated. Extraversion polygenic scores were found to *positively* predict bipolar disorder and psychological distress[60,61], but also to *negatively* predict anxiety/depression or related traits. The latter is more in line with the negative phenotypic correlation between extraversion and anxiety, depression and bipolar disorder[61,62]. However, extraversion is related to a manic prone illness course in bipolar disorder[62]. Again, sample size seems to influence the variation in results.

Sub-type analysis

In addition to shared genetic risk across diagnostic classes, heterogeneity within diagnostic classes is well-recognised in psychiatry. It is appealing to attempt to use genome-wide association data to explore if genetic heterogeneity underpins the phenotypic heterogeneity. Currently, such analyses are often limited by sample size (**Table 3**, **Supplementary Table 4**), and to date most applications have been on schizophrenia subtypes. Limited sample sizes have prohibited GREML analyses, to date. We anticipate that sub-type analyses will become more common in the future and GPRS provides a mechanism to use the power of a larger discovery sample recorded only for case-control status to probe differences between subtypes recorded in smaller samples of the same or different disorder. Extending such analyses to larger samples requires consistent phenotypes across samples, which may be a problem. For example, many subtypes of schizophrenia have been proposed based on cognitive deficit, symptom profiles or treatment resistance. But, to date, GPRS results for schizophrenia subtyping do not show a clear pattern (**Table 3**), which may reflect, in part, heterogeneity of the discovery sample. The most interpretable application of sub-type analysis is between one disorder (say schizophrenia since it currently has the most powerful discovery sample) and sub-types of another disorder. Applications between one disorder and subtypes of the same disorder may be more difficult to interpret. For example, if the PGC schizophrenia sample is used as a discovery sample to compare polygenic risk scores in independent samples of cognitive deficit vs cognitive normal cases or between clozapine treated (usually treatment resistant) and non-clozapine cases the detailed interpretation depends on the proportion of these subtypes in the discovery sample, which is likely unknown. Nonetheless, some interpretation is possible and, for example, GPRS scores from an ADHD discovery samples were higher in target samples of ADHD cases with conduct disorders vs ADHD samples without conduct disorder[63], implying some genetic basis for differences in the etiology of these classes. Interestingly, application of ADHD GPRS in a population sample of children measured at 7 and 10 years for ADHD related traits provided empirical evidence for the hypothesis that hat ADHD represents the extreme end of traits present in the general population[64].

Discussion

Why are common loci of small effect important?

Other than for schizophrenia, the number of genome-wide significant DNA variants identified for psychiatric disorders or related traits is few, to date. The success for schizophrenia is largely explained by greater sample size which was achieved by combining data across >50 studies[65]. Indeed, no single locus had been robustly associated with schizophrenia when sample sizes were similar to those currently available for many other psychiatric disorders. The trajectory of GWAS discovery for schizophrenia, which increased from 1 locus[66,67] to 7 (ref [68]) to 22 (ref [69]) to 62 (ref [70]) to >100 (ref[65]) as the number of cases increased from ~3K to 36K, is not dissimilar to that of other (non-psychiatric) diseases[71]. Whereas, the first 3K cases identified only 1 risk variant, the last 3K cases added to make the total of 36K cases identified ~40 additional loci[65]. This success was predicted from polygenic analyses applied to the data sets that found only 1 locus[8,39]. The same polygenic analyses applied to other psychiatric disorders imply that common SNPs explain a significant proportion of variance implying that current sample sizes are underpowered to detect the effect sizes that exist in nature, and that more individually associated loci for these disorders will be identified as sample size increases. However, collection of larger samples is time consuming and expensive, so why is it important to identify common loci of small effect?

First, although verified GWAS effects are usually small individually, their cumulative effect is not. Second, there is evidence that loci found to harbour common alleles of small effect for schizophrenia are also enriched for rare mutations of larger effect in whole exome sequencing studies in schizophrenia (e.g. voltage-gated calcium channel genes)[72]. Convergent results from GWAS and sequencing can help to prioritise genes for follow-up studies[73]. Third, small effect size may partly reflect the heterogeneity of the diagnostic construct that is imposed from a diagnostic system based on self-report and clinical observation. As sample sizes and genomic technology improve, the genomics era has the potential to identify more biologically based diagnostic constructs for which effect sizes may be larger. Fourth, there are now many examples of diseases for which GWAS hits are for known drug targets[74,75] or identify relevant biology[76,77]. For example, genes identified through GWAS associated with variation in LDL levels are the targets of statins[78] and those associated with rheumatoid arthritis are the targets of known drugs that are effective therapies for this disease[79]. Similar insights may be forthcoming in psychiatric disorders, because identified loci for schizophrenia include known targets of existing antipsychotics[65]. These examples indicate that although GWAS loci have small effect sizes, they may help identify targets for novel therapeutics[78], or may identify existing drugs that can be repurposed for treatment of diseases that they were not initially developed to treat[75,80]. For these reasons identification of common variants of small effect is a worthy goal and polygenic methods provide strong guidance based on currently available data that increasing sample size will identify more associated variants in future. To maximise the value of new sample collections they should be accompanied by more detailed clinical data.

Implications for nosology

The results from the cross disorder analyses are an important outcome of polygenic methods, providing direct empirical evidence for genetic relationships between disorders. These results will contribute to nosology. The genome-wide era provides a new paradigm to explore the genetic relationship between disorders. In the pre-genomics era genetic relationships between disorders could only be determined by collection of large cohorts of families measured for the two disorders. Low population risk, variable age of onset, ascertainment biases and confounding with family environment make such data difficult to collect. For examples where the genetic relationship between disorders has been investigated through twin and family studies, the results generally converge with the latest results obtained with polygenic methods (see citations in[41]). This supports further use of genome-wide SNPs to explore the genetic relationship between case-control samples collected independently for pairs of disorders. Another approach to explore the genetic overlap between disorders is the conditional false discovery rate (cFDR) method[81] in which the search space for associated SNPs in the target sample is limited to SNPs associated to some threshold in the discovery sample. In this way, more true-positive associated SNPs surpass the stringent level of association significance. In contrast to GPRS and bivariate GREML methods, cFDR is agnostic to direction of effect and so considers a more general **pleiotropy** in which the same SNPs but different risk alleles can be identified. Lastly, as discussed above and below, increased sample size accompanied by consistent clinical data and advances in genomic technology have the potential to add knowledge to both shared genetic factors across disorders and heterogeneity within disorders to create more biologically valid diagnostic constructs.

Where to find the still-missing heritability

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}. However, the “still-missing” heritability, i.e. the difference between h^2_{SNP} and h^2 remains substantial for psychiatric disorders and, indeed, for most other complex traits (at least half) is still missing. 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. In human populations, part of the still-missing heritability may simply reflect overestimation of h^2 since typical human family 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[82,83]. The difference between estimates of h^2 from family data and the “true” h^2 has been termed “phantom heritability”[84] 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[85,86] and that presence of important **epistasis** and small **epistatic variance** are not inconsistent[87]. Empirical support comes from the study of gene transcription levels for which significant, replicated epistatic effects have been identified but these explain only one-tenth of the variance compared to additive variance[88].

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[89]. For this reason, studies of candidate GxE interactions in psychiatry have generally lacked replication and the field is plagued by publication bias towards studies with positive results[90]. Polygenic risk scores analyses provide a novel paradigm to quantify GxE[91,9293].

Part of the still-missing heritability must reflect genomic variants not well tagged by SNPs[16,21]. Since the SNPs on SNP chips are chosen because both their alleles are common they cannot be in high r^2 linkage disequilibrium with a causal variant with one rare allele. A very large number of rare variants are needed to explain the still-missing heritability, since such variants individually explain a very small proportion of the variance. 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 the relative importance of small structural variants to genomic variation is currently not well documented and since recurrent tandem repeat polymorphisms are known to modulate a range of biological functions[94,95] these may represent an example of an important, but as yet unprobed, source of disease associated variation.

Disorder heterogeneity is a possible explanation for still-missing heritability of particular relevance to psychiatric disorders. One aspect of disorder heterogeneity may be reflected by lower estimates of heritabilities of psychiatric disorders from large national registries than from clinically ascertained cohorts. For example, estimates of h^2 for schizophrenia and bipolar disorder were 0.64 (95% CI 0.62-0.68) and 0.59 (0.56-0.62) from the Swedish national data[96], 0.67 (0.64–0.71), 0.62 (0.58–0.65) estimated from reported summary statistics of Danish national data[97], compared to estimates from meta-analysis of clinically ascertained studies[98,99] of 0.81 (0.73-0.90) and 0.85 (0.73-0.93). The lower estimates from national data may reflect, in part, differences in diagnostic criteria that may be more relevant to the large samples brought together for genome-wide genotyping, whereas careful and consistent diagnostic practice is likely to be used in the clinical samples ascertained for estimation of heritability. However, another aspect of diagnostic heterogeneity may be that biologically different disorders are labelled the same given the clinically available symptom data. For illustrative purposes, consider a non-psychiatric paradigm. Crohn's disease and ulcerative colitis are both forms of inflammatory bowel disease (IBD) and based on patient symptoms and clinical observation it is difficult to discriminate between them. It is only in the last forty-odd years, with the advent of colonoscopy, that differential diagnosis has become possible. GWAS have identified 163 IBD loci, the vast majority of which have odds ratio in the same direction for both disorders[100]. Despite the strong common biological mechanisms, many of the risk alleles have significantly different effect sizes between the disorders, and it is notable that two risk alleles for Crohn's Disease (in *PTPN22* and *NOD2*) are significantly protective for ulcerative colitis. In other words these loci would not be identified or would be identified with reduced odds ratio in association analysis of IBD. The parallels with psychiatric disorders are clear (although the differences may be more subtle), currently we may not have the phenotypic benchmarks to allow subtype distinction of disorders and hence variants that differentiate between subtypes may be obscured. 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[101,102] or K-ras mutations in colorectal cancer and EGFR mutations in lung cancer, reviewed in[103]), however other branches of medicine are able to supply measures of phenotypic heterogeneity in the tissue of relevance for mapping onto the genetic heterogeneity. Lack of access to brain tissue will make progress slower in psychiatry. It is well recognised that affected family members tend to have more similar symptom profiles than with other affected individuals[104-106], implying that if association analyses were limited to a subset of individuals with similar symptom profiles effect sizes of individual variants may be higher. In other words, the current disease classification might obscure different subtypes and identifying the subtypes that currently cannot be differentiated is an important goal of psychiatric genomics. Especially of relevance in child psychiatry is the heterogeneity in course of psychiatric disorders from childhood into adolescence and then adulthood. One apparent example is ADHD for which symptoms can persist into adulthood, but can also decline[107]. It is yet unknown whether this difference in course represents genetically heterogeneous sub-types. Another issue is the transition of childhood symptoms into a range of other adulthood disorders, recently studied in the context of SNPs suggestively associated with adulthood mood and psychotic disorders investigated in childhood ADHD and ASD and internalizing symptoms at age 3 (ref[49]). In summary, the most tangible way forward to gain a more complete picture of still-missing heritability are large samples informative for G, E and clinical symptoms.

Recommendations for polygenic analyses

In this review we have considered the polygenic methods most commonly applied to psychiatric disorders and traits, namely GPRS and GREML. The GPRS results interpreted through simulation[8] and theory[34] generate estimates of h^2_{SNP} consistent with those from GREML. That three different methods generate consistent results provides some support for the robustness of the estimates, although of course they are detecting the same underlying signal and make some similar assumptions. GREML estimates of h^2_{SNP} and SNP-correlations from genome-wide SNP data considered jointly in a single analysis are statistically optimal, are robust to perturbations in underlying assumptions and explicit test for inflation by artefacts. The robustness of the method to underlying assumptions has been tested in detail[108,109]. As sample sizes increases even more interesting partitioning of variance based on annotation of SNPs will become possible. However, application of GREML requires access to genome-wide genotypes for all samples, which is not always possible, whereas GPRS requires genome-wide genotypes only for the target sample. Moreover, GREML is orders of magnitude more demanding in computing resources than GPRS. The validity of the approximate method[34] needs to be tested further to determine if any assumptions impact results compared to those calculated from the statistically more complete methods. We recommend application of the GPRS plus the Dudbridge[34] approximation alongside GREML estimates so that this can be fully evaluated. Application of the Dudbridge approximation to published GPRS results is difficult because not all the needed input parameters have been reported. A potential pitfall is that, currently, it is not possible to determine if there is overlap between discovery and target sample when only summary statistics are available for the discovery sample and overlap would serve to inflate results[18]. However, as a rule of thumb if application of GPRS shows no significant variance explained then there is little point in bothering with more refined analyses. However, if GPRS analyses provide evidence for an important polygenic component then GREML analyses use data in the optimal way. For example, integration of functional annotation into GPRS methods is limited because arbitrary decisions are made about which SNPs out of sets of correlated SNPs are retained in the analysis. In GREML analyses the correlation structure between SNPs is accounted for, and hence the data drives how variance is attributed to different functional classes[110].

Current and future value of genomic risk predictors

GPRSs and individual estimated genetic values (a by-product of GREML analysis, **Box 2** step 7) are risk prediction scores for individuals. These are currently not of diagnostic value and indeed a genetic predictor alone will always have limited predictive value when the heritability is less than 1 (ref[111]). Even as sample sizes increase their utility will be limited to identification of high-risk strata that may contain the majority of individuals who are or become affected, even though the majority of individuals in the high-risk strata may not be affected (i.e., high sensitivity, low specificity)[32,112]. However, predictive ability will increase if non-genetic risk factors are combined with the genetic predictors. Moreover, genetic studies may lead to the identification of other biomarkers, such as proteomic biomarkers[113], through discovery of novel pathways. The challenge in psychiatry is not the classification into cases vs controls, but into treatment relevant subsets amongst individuals presenting in prodromal phase of their disorder trajectory[114]. Despite these challenges predictors may be a tangible outcome of the genomics era, as understanding of biological mechansims are not needed for classifiers to have clinical utility[114].

Conclusions

The genomics era has provided the empirical evidence that complex genetic diseases and disorders are indeed complex. The complexity can seem overwhelming but the genomic data has provided some traction. Only time will tell what the knowledge of the 100+ loci detected to date for schizophrenia[65] will deliver in terms of prevention, diagnosis, prognosis and treatment option and whether the number of risk loci identified for other disorders will increase as predicted with increasing sample size. However, we conclude that polygenic analyses will contribute further to our understanding of complex genetic traits as sample sizes increase and as sample resources become richer in phenotypic descriptors, both in terms of clinical symptoms and of non-genetic risk factors.

Acknowledgements

We acknowledge financial support from the Australian Research Council (FT0991360, DE130100614), the Australian National Health and Medical Research Council (613602, 1047956, 1011506), the Netherlands Organisation for Health Research and Development (91210020)

URLs

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink
GREML: http://www.complextraitgenomics.com/software/gcta//
GREML power calculator: http://spark.rstudio.com/ctgg/gctaPower/
ABC tool to convert Nagelkerke's R2 to variance in liability explained:
http://www.complextraitgenomics.com/software/
Dudbridge approximation: sites.google.com/site/ fdudbridge/software/

References

1. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-76 (2010).
2. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24 (2012).
3. Chanock, S.J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
5. Neale, B.M. *et al.* Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 884-97 (2010).
6. Anney, R. *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072-82 (2010).
7. Sklar, P. *et al.* Whole-genome association study of bipolar disorder. *Mol Psychiatry* **13**, 558-69 (2008).

8. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).

9. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-7 (2009).

10. Sullivan, P.F. *et al.* Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatry* **14**, 359-75 (2009).

11. Gottesman, II & Shields, J. A polygenic theory of schizophrenia. *Proc Natl Acad Sci U S A* **58**, 199-205 (1967).

12. Thapar, A. & Harold, G. Why is there such a mismatch between traditional heritability estimates and molecular genetic findings for behavioural traits? . *Journal of Child Psychology & Psychiatry* **This volume**(2014).

13. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847-56 (2000).

14. So, H.C., Gui, A.H., Cherny, S.S. & Sham, P.C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* **35**, 310-7 (2011).

15. Sham, P. *Statistics in human genetics*, (Arnold, London, 1998).

16. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).

17. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* **36**, 214-24 (2012).

18. Wray, N.R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* **14**, 507-15 (2013).

19. Kooperberg, C., LeBlanc, M. & Obenchain, V. Risk prediction using genome-wide association studies. *Genet Epidemiol* **34**, 643-52 (2010).

20. Abraham, G., Kowalczyk, A., Zobel, J. & Inouye, M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet Epidemiol* **37**, 184-95 (2013).

21. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).

22. Habier, D., Fernando, R.L. & Dekkers, J.C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97 (2007).

23. VanRaden, P.M. Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414-23 (2008).

24. Goddard, M.E. Genomic Selection: predicion of accuracy and maximisation of long term response. *Genetica* **136**, 245-257 (2009).

25. Fernando, R.L. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production,* **26**, 329-336 (1998).

26. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294-305 (2011).

27. Benjamin, D.J. *et al.* The genetic architecture of economic and political preferences. *Proc Natl Acad Sci U S A* **109**, 8026-31 (2012).

28. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

29. Wray, N.R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res Hum Genet* **8**, 87-94 (2005).

30. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2 (2012).
31. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807-12 (2011).
32. So, H.C., Kwan, J.S., Cherny, S.S. & Sham, P.C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am J Hum Genet* **88**, 548-65 (2011).
33. So, H.C., Li, M. & Sham, P.C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet Epidemiol* **35**, 447-56 (2011).
34. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348 (2013).
35. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **In press**(2014).
36. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-41 (2008).
37. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232-6 (2008).
38. Sullivan, P. Don't give up on GWAS. *Mol Psychiatry* **17**, 2-3 (2012).
39. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44**, 247-50 (2012).
40. Lee, S.H. *et al.* Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet* **22**, 832-41 (2013).
41. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).
42. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism* **3**, 9 (2012).
43. St Pourcain, B. *et al.* Common variation contributes to the genetic architecture of social communication traits. *Mol Autism* **4**, 34 (2013).
44. Davis, L.K. *et al.* Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* **9**, e1003864 (2013).
45. Trzaskowski, M., Dale, P.S. & Plomin, R. No genetic influence for childhood behavior problems from DNA analysis. *J Am Acad Child Adolesc Psychiatry* **52**, 1048-1056 e3 (2013).
46. Trzaskowski, M. *et al.* First genome-wide association study on anxiety-related behaviours in childhood. *PLoS One* **8**, e58676 (2013).
47. Viding, E. *et al.* Genetics of callous-unemotional behavior in children. *PLoS One* **8**, e65789 (2013).
48. Vinkhuyzen, A.A. *et al.* Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Transl Psychiatry* **2**, e102 (2012).
49. Benke, K. & al, e. A genome-wide association meta-analysis of Preschool Internalizing Problems. *Journal of the American Academy of Child and Adolescent Psychiatry* (In press).
50. Jiang, Y.H. *et al.* Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. *Am J Hum Genet* (2013).

51. de Candia, T.R. *et al.* Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. *Am J Hum Genet* **93**, 463-70 (2013).

52. Yang, L. *et al.* Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. *Am J Med Genet B Neuropsychiatr Genet* **162B**, 419-30 (2013).

53. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-9 (2013).

54. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167**, 1357-63 (2010).

55. Larsson, H. *et al.* Risk of bipolar disorder and schizophrenia in relatives of people with attention-deficit hyperactivity disorder. *Br J Psychiatry* **203**, 103-6 (2013).

56. Rommelse, N.N., Franke, B., Geurts, H.M., Hartman, C.A. & Buitelaar, J.K. Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. *Eur Child Adolesc Psychiatry* **19**, 281-95 (2010).

57. Sullivan, P.F. *et al.* Family history of schizophrenia and bipolar disorder as risk factors for autism *Archives of General Psychiatry* **69**, 1099-1103 (2012).

58. Vorstman, J.A. *et al.* No evidence that common genetic risk variation is shared between schizophrenia and autism. *Am J Med Genet B Neuropsychiatr Genet* **162B**, 55-60 (2013).

59. Hamshere, M.L. *et al.* Shared polygenic contribution between childhood attention-deficit hyperactivity disorder and adult schizophrenia. *Br J Psychiatry* **203**, 107-11 (2013).

60. Luciano, M. *et al.* Genome-wide association uncovers shared genetic effects among personality traits and mood states. *Am J Med Genet B Neuropsychiatr Genet* **159B**, 684-95 (2012).

61. Middeldorp, C.M. *et al.* The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data. *Transl Psychiatry* **1**, e50 (2011).

62. Barnett, J.H. *et al.* Personality and bipolar disorder: dissecting state and trait associations between mood and personality. *Psychol Med* **41**, 1593-604 (2011).

63. Hamshere, M.L. *et al.* High loading of polygenic risk for ADHD in children with comorbid aggression. *Am J Psychiatry* **170**, 909-16 (2013).

64. Martin, J., Hamshere, M.L., Stergiakouli, E., O'Donovan, M.C. & Thapar, A. Genetic risk for attention-deficit/hyperactivity disorder contributes to neurodevelopmental traits in the general population. *Biological Psychiatry* **In press**(2014).

65. Ripke, S. Psychiatric Genomics Consortium quadruples schizophrenia GWAS sample-size to 35,000 cases and 47,000 controls. in *XXIst World Congress of Psychiatric Genetics: Redefining mental illness through genetics* (Boston, Massachusetts, 2013).

66. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

67. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757 (2009).

68. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**, 969-76 (2011).

69. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-9 (2013).
70. Anderson-Schmidt, H. *et al.* Selected rapporteur summaries from the XX World Congress of Psychiatric Genetics, Hamburg, Germany, October 14-18, 2012. *Am J Med Genet B Neuropsychiatr Genet* **162B**, 96-121 (2013).
71. Kim, Y.J., Zerwas, S., Trace, S.E. & Sullivan, P.F. Schizophrenia Genetics: Where Next? *Schizophrenia Bulletin* **37**, 456-463 (2011).
72. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).
73. Gratten, J., Visscher, P.M., Mowry, B.J. & Wray, N.R. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet* **45**, 234-8 (2013).
74. Plenge, R.M., Scolnick, E.M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* **12**, 581-94 (2013).
75. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* **30**, 317-20 (2012).
76. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).
77. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
78. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).
79. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* (2013).
80. Manolio, T.A. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* **14**, 549-58 (2013).
81. Schork, A.J. *et al.* All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet* **9**, e1003449 (2013).
82. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).
83. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139-49 (2013).
84. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193-8 (2012).
85. Stringer, S., Derks, E.M., Kahn, R.S., Hill, W.G. & Wray, N.R. Assumptions and properties of limiting pathway models for analysis of epistasis in complex traits. *PLoS One* **8**, e68913 (2013).
86. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
87. Mackay, T.F. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet* **15**, 22-33 (2014).
88. Hemani, G. *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* (2014).
89. Dunn, E.C. *et al.* Research review: gene-environment interaction research in youth depression - a systematic review with recommendations for future research. *J Child Psychol Psychiatry* **52**, 1223-38 (2011).
90. Duncan, L.E. & Keller, M.C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am J Psychiatry* **168**, 1041-9 (2011).

91. McGrath, J.J., Mortensen, P.B., Visscher, P.M. & Wray, N.R. Where GWAS and epidemiology meet: opportunities for the simultaneous study of genetic and environmental risk factors in schizophrenia. *Schizophr Bull* **39**, 955-9 (2013).
92. Plomin, R. Commentary: missing heritability, polygenic scores, and gene-environment correlation. *J Child Psychol Psychiatry* **54**, 1147-9 (2013).
93. Iyegbe, C., Campbell, D., Butler, A., Ajnakina, O. & Sham, P. The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for GxE research. *Soc Psychiatry Psychiatr Epidemiol* **49**, 169-82 (2014).
94. Hannan, A.J. TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease. *Discov Med* **10**, 314-21 (2010).
95. Hannan, A.J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet* **26**, 59-65 (2010).
96. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234-239 (2009).
97. Service, S.K. *et al.* A genome-wide meta-analysis of association studies of Cloninger's Temperament Scales. *Transl Psychiatry* **2**, e116 (2012).
98. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait - Evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187-1192 (2003).
99. McGuffin, P. *et al.* The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry* **60**, 497-502 (2003).
100. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
101. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
102. Slamon, D.J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-82 (1987).
103. Ferraldeschi, R. & Newman, W.G. Pharmacogenetics and pharmacogenomics: a clinical reality. *Ann Clin Biochem* **48**, 410-7 (2011).
104. Schreier, A., Hofler, M., Wittchen, H.U. & Lieb, R. Clinical characteristics of major depressive disorder run in families--a community study of 933 mothers and their children. *J Psychiatr Res* **40**, 283-92 (2006).
105. Kendler, K.S. *et al.* Resemblance of psychotic symptoms and syndromes in affected sibling pairs from the Irish Study of High-Density Schizophrenia Families: evidence for possible etiologic heterogeneity. *Am J Psychiatry* **154**, 191-8 (1997).
106. Lieb, R., Isensee, B., Hofler, M. & Wittchen, H.U. Parental depression and depression in offspring: evidence for familial characteristics and subtypes? *J Psychiatr Res* **36**, 237-46 (2002).
107. Faraone, S.V., Biederman, J. & Mick, E. The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychol Med* **36**, 159-65 (2006).
108. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-21 (2012).
109. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum Genet* **131**, 1655-64 (2012).

110. Gusev, A. *et al.* Quantifying missing heritability from coding variation in schizophrenia in *XXIst World Congress of Psychiatric Genetics: Redefining mental illness through genetics* (Boston, Massachusetts, 2013).

111. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS genetics* **6**, e1000864 (2010).

112. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* **45**, 400-5, 405e1-3 (2013).

113. Sokolowska, I. *et al.* The potential of biomarkers in psychiatry: focus on proteomics. *J Neural Transm* (2013).

114. Kapur, S., Phillips, A.G. & Insel, T.R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry* **17**, 1174-9 (2012).

115. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

116. Collins, A.L. *et al.* Identifying bipolar disorder susceptibility loci in a densely affected pedigree. *Mol Psychiatry* (2012).

117. Whalley, H.C. *et al.* The influence of polygenic risk for bipolar disorder on neural activation assessed using fMRI. *Transl Psychiatry* **2**, e130 (2012).

118. Demirkan, A. *et al.* Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Mol Psychiatry* **16**, 773-83 (2011).

119. Otowa, T. *et al.* Meta-analysis of genome-wide association studies for panic disorder in the Japanese population. *Transl Psychiatry* **2**, e186 (2012).

120. Derks, E.M., Vorstman, J.A., Ripke, S., Kahn, R.S. & Ophoff, R.A. Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis. *PLoS One* **7**, e37852 (2012).

121. Ikeda, M. *et al.* Genome-wide association study of schizophrenia in a Japanese population. *Biol Psychiatry* **69**, 472-8 (2011).

122. Levinson, D.F. *et al.* Genome-wide association study of multiplex schizophrenia pedigrees. *Am J Psychiatry* **169**, 963-73 (2012).

123. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* (2013).

124. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum Mol Genet* **21**, 4781-92 (2012).

125. Vrieze, S.I., McGue, M. & Iacono, W.G. The interplay of genes and adolescent development in substance use disorders: leveraging findings from GWAS meta-analyses to test developmental hypotheses about nicotine consumption. *Hum Genet* **131**, 791-801 (2012).

126. Whalley, H.C. *et al.* Polygenic risk and white matter integrity in individuals at high risk of mood disorder. *Biol Psychiatry* **74**, 280-6 (2013).

127. Holmes, A.J. *et al.* Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. *J Neurosci* **32**, 18087-100 (2012).

128. Bigdeli, T.B. *et al.* Molecular Validation of the Schizophrenia Spectrum. *Schizophr Bull* (2013).

129. Fanous, A.H. *et al.* Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *Am J Psychiatry* **169**, 1309-17 (2012).

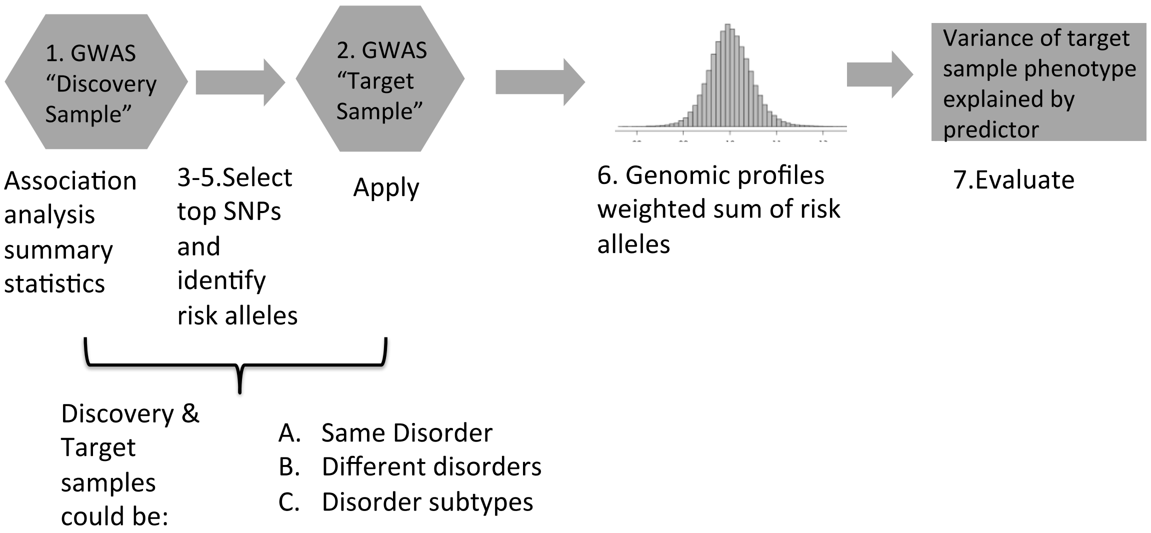
130. Hamshere, M.L. *et al.* Polygenic dissection of the bipolar phenotype. *Br J Psychiatry* **198**, 284-8 (2011).

Glossary

Clumping – selection of SNPs based on association p-value and LD threshold between SNPs to generate a SNP set that includes the most associated SNP within LD regions.
Complex genetic trait - a trait or disease that tends to "run in families" but shows no clear pattern of inheritance and so is likely underpinned by multiple genetic and non-genetic factors
Conventionally unrelated - Individuals from that are not closely related, for example more distantly related than 2nd cousins.
Epistasis - nonlinear interactions between segregating loci; when the phenotypic effect of the genotype at one locus depends on the genotype at another locus
Epistatic variance – the variance partitioned out from the total genetic variance that is orthogonal to additive genetic variance, i.e., additive variance is the variance attributable to average effects and the epistatic variance is the variance attributable to deviations from average effects. Hence, epistasis contributes to both average effects and to deviations from additive effects and the presence of epistasis and small epistatic variance are not inconsistent.
GREML- genomic-relationship-matrix restricted maximum likelihood; a method to estimate the variances of random effects from a mixed linear model in which the correlation structure between the genetic random effects is defined by the genomic relationship matrix calculated from SNPs.
Linkage disequilibrium (LD)- Two alleles at different loci that occur together on a chromosome more often than would be predicted by random chance.
Pleiotropy- the phenotypic effect of a genetic variant on more than one trait
Polygenic – a genetic architecture of "many" genetic variants and includes risk variants that are both common and rare in the population.
Population stratification - Structure within the sample due to differences in genetic ancestry among samples.
Profile Score – a weighted sum of the number of risk alleles carried by an individual. The risk alleles and their effect sizes (the weights) are calculated from an independent sample.

Box 1 Mini Guide to Method: Genomic Profile Risk Scoring
Method:

1. Identify Discovery sample with genome-wide association analysis summary statistics
2. Identify Target sample with genome-wide genotypes. The Target sample should not include individuals closely related to those in the Discovery sample. Results can be inflated if there is overlap between samples.
3. Determine the list of SNPs in common between Discovery and Target samples
4. Construct a clumped SNP list: association p-value informed removal of correlated SNPs, e.g. LD threshold of $r^2 < 0.2$ across 500 kb. (e.g.,in the program PLINK[115]: –clump-p1 1–clump-p2 1–clump-r2 0.2–clump-kb 500)
5. Limit SNP list to those with association p-value less than a defined threshold (often several thresholds are considered, i.e., <0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3 etc).
6. Generate genomic profile scores in the target sample: e.g., sum of risk alleles weighted by Discovery sample log(odds ratio). (e.g., in PLINK: –score)
7. Regression analysis: y= phenotype, x = profile score. Compare variance explained from the full model (with x) compared to a reduced model (covariates only). Check the sign of the regression coefficient to determine if the relationship between y and x is in the expected direction.


1. GWAS "Discovery Sample"
2. GWAS "Target Sample"
Variance of target sample phenotype explained by predictor

Association analysis summary statistics

3-5.Select top SNPs and identify risk alleles

Apply

6. Genomic profiles weighted sum of risk alleles

7.Evaluate

Discovery & Target samples could be:

A. Same Disorder
B. Different disorders
C. Disorder subtypes

Outcomes:

1) Measure of association between Discovery and Target sample (R^2, Nagelkerke's R^2, area under the receiver operating curve, proportion of variance explained on liability scale, see[17])
2) Genomic profile risk score values for each individual in the Target sample. These can be used in future experimental design, for example, imaging studies that compare those with high and low polygenic risk score.

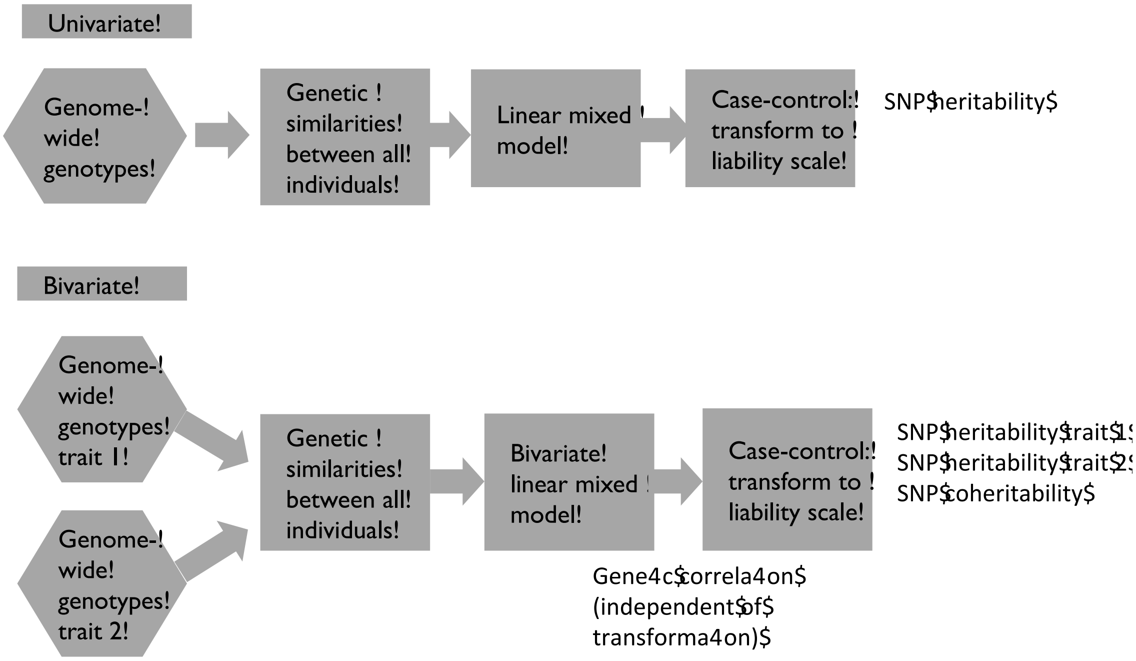
Sample size considerations:

1) To maximise the test statistic of association between Discovery and Target, these samples should be of equal size[32]. Under realistic assumptions there is sufficient power to detect a significant proportion of variance explained when the Target sample is ~2000 individuals. A useful rule of thumb is then to make Discovery and Target samples of equal size until the Target sample is ~2000 individuals and then allocate additional samples to the Discovery sample to maximise the accuracy of the GPRS for individuals[32].
2) The p-value threshold (step 5) that maximises the variance explained in the Target sample depends on sample size of the Discovery sample and the underlying unknown genetic architecture. In discovery samples that are underpowered for a GWAS (for example, those that identify few genome-wide significant associations) it is not uncommon to find that predictive

ability in the target sample is maximised when a majority or all SNPs are included in the profiling SNP list. However, simulations (see Figure S8 in[8]) show that as discovery sample size increases the change in the pattern of the predictive ability of the SNP set at different p-value thresholds reflects the underlying genetic architecture (i.e., number, frequency spectrum, and effect size distribution of truly associated variants). The emerging empirical data confirms the simulations: lower p-value thresholds maximise predictive ability in the target sample as discovery samples increase[69], reflecting that with larger sample sizes true positives become more enriched in the SNP sets with lower p-value thresholds.

Box 2: Mini Guide to method: Estimating variance explained by all SNPs
Method:

1. Identify data sets with genome-wide SNP genotypes. For bivariate analyses the same individuals or different sets of individuals can be measured for the two traits.
2. Apply more stringent QC than for standard GWAS analysis
3. Calculate the genome relationship matrix (GRM) – a matrix of genome-wide similarities between all pairs of individuals calculated from the genome-wide SNPs. (In the program GCTA[28]: --make-grm). Multiple GRM can be made based on SNP annotation to allow partitioning of variance.
4. Exclude one of each pair of individuals who are more related than chosen threshold – usually no more related than second cousins, so that estimates reflect the signal tagged by common variants through population level LD (e.g., in GCTA: --grm-cutoff 0.025)
5. Estimate variance attributable to SNPs via residual maximum likelihood (REML) analysis from a linear mixed model with covariates (e.g., in GCTA: --reml).
6. For case-control analysis transform the result to the liability scale (e.g., in GCTA: –prevalence)
7. Best Linear Unbiased Predictions of total genetic values for individuals on the untransformed scale can be derived (but not yet implemented in GCTA).


Univariate!
Genome-! wide! genotypes!
Genetic ! similarities! between all! individuals!
Linear mixed ! model!
Case-control:! transform to ! liability scale!
SNP$heritability$
Bivariate!
Genome-! wide! genotypes! trait 1!
Genome-! wide! genotypes! trait 2!
Genetic ! similarities! between all! individuals!
Bivariate! linear mixed ! model!
Case-control:! transform to ! liability scale!
SNP$heritability$trait$1
SNP$heritability$trait$2
SNP$coheritability$
Gene4c$correla4on$ (independent$of$ transforma4on)$

Outcomes:

From univariate analysis we estimate the proportion of variance attributable to SNPs or SNP-heritability (sometimes called chip-heritability), h^2_{SNP}. From bivariate analyses we estimate SNP-heritabilities for each of the two traits (h^2_{SNP-1}, h^2_{SNP-2}), the SNP-genetic correlation between them ($r_{g\text{-}SNP}$) and the coheritability between them $r_{g\text{-}SNP}h_{SNP\text{-}1}h_{SNP\text{-}2.}$ N.B. The correlation is independent of scale (i.e., is the same before and after transformation). If multiple GRM are fitted (step 3) then variance is partitioned according to SNP annotation (e.g. chromosome, frequency, function).

Sample size considerations:

Outcome estimates are unbiased, therefore as sample sizes increase the standard errors of the estimates decrease, but the estimates should not change given the bounds indicated by the standard errors. This assumes that all individuals are samples from the same idealised genetically homogenous population, and excluding genotyping artefacts. If estimates change more than expected given the standard error then this assumption may be violated.

It is not ideal to undertake meta-analysis of estimates from individual samples rather than analysing the total sample together. The s.e of h^2_{SNP} from a total sample of 10,000 but meta-analysed from five estimates each of 2000 is 0.072 compared to the s.e of 0.032 when all 10,000 samples are analysed together. This is because the genetic relationships between individuals in different sub-samples are not used in meta-analysis.

Box 3. The relationship between GPRS and GREML

GPRS and GREML can be applied to the same data sets and both can provide evidence for a polygenic contribution to the trait or a shared polygenic relationship between traits that is tagged by the common SNPs. Necessarily the methods must be tapping into the same signal provided by the data. In the first application of GPRS in the study of the International Schizophrenia Consortium (ISC), extensive simulations were undertaken in order to understand the likely underlying genetic architecture that could generate the observed results. The simulations showed that a range of underlying genetic architectures (in terms of number, frequency and effect size of causal variants) could have generated the observed GPRS results in which the ISC Discovery sample (3322 cases, 3587 controls) generated a Nagelkerke's R^2 of 0.032 p-value of $2x10^{-28}$ in the Molecular Genetics of Schizophrenia (MGS, 2687 cases, 2656 controls) target sample based on SNPs with p-value threshold 0.5 out of 74062 LD-pruned SNPs. However, all simulation genetic architectures that were consistent with the empirical results pointed to a h_{SNP}^2 of 0.34. Application of GREML to the ISC data generated a direct estimate of h_{SNP}^2 = 0.33 (95% CI 0.24-0.42) Supplementary Table 2 in ref[39]), reducing to h_{SNP}^2 = 0.27 (95% CI 0.21-0.33) after stringent QC, designed to reduce the chances that the reported estimate is inflated by artefacts such as population stratification. These results demonstrate the relationship between GPRS and GREML via simulation. Using regression theory and, for case-control studies the liability threshold model, Dudbridge[34] provided the theoretical framework to directly estimate h_{SNP}^2 from same disorder applications of GPRS and $r_{g\text{-}SNP}$ from cross-disorder applications of GPRS. For example, using the ISC sample characteristics as described above, his R code calculator "estimateVg2FromP" (p=2e-28, n1=3322+3587, nsnp=74062, n2=2687+2656, vg1=0, corr=1, plower=0, pupper=0.5, weighted=T, binary=T, prevalence1=.01, prevalence2=.01, sampling1=3322/(3322+3587), sampling2=2687/(2687+2656), lambdaS1=NA, lambdaS2=NA, nullfraction=0, shrinkage=F, logrisk=F) generates an estimate of h_{SNP}^2 = 0.287 (95% CI 0.236 - 0.337).

We applied the "estimateCorrFromP" calculator to the GPRS results from the Psychiatric Genomics Consortium Cross Disorder Group (PGC-CDG) analyses[53] (their Figure 3 and Table S5, see Table S6). The estimated SNP-correlation is compared to the direct GREML estimate of the SNP-correlation (**Figure 3**) and shows good agreement, particularly in terms of benchmarking high, medium or low correlation. In this example, the application of the Dudbridge approximation is optimised as an input to the calculation of h_{SNP}^2 for which the GREML estimates are used. The validity of the approximate method[34] needs to be tested further to determine if any assumptions impact results compared to those calculated from the statistically more complete methods.

Table 1: Studies using GPRS with discovery and target samples of the same trait (univariate analysis). Above: adult psychiatric disorders, below psychiatric disorders usually diagnosed during childhood.

Phenotype	**Reference**
Alcohol dependence	[65]
Bipolar disorder	[116,117]
Cloninger's temperament scales: harm avoidance, novelty seeking, reward dependence, persistence	[97]
Major depressive disorder	[118]
Neuroticism and extraversion	[60]
Panic disorder	[119]
Schizophrenia	[8,120-123]
ADHD	[63]
Autism spectrum disorders	[124]
Behavioral disinhibition, alcohol use, drug use, nicotine use/dependence	[125]

See Table S2 for more details

Table 2: Overlap in genetic risk factors between two different disorders/traits, GPRS studies

Discovery: Target Phenotypes	**Evidence for genetic overlap**	**Reference**
BD: BD with family history, brain activation during sentence completion test	BD activation in anterior cingulated cortex and the right amygdala across case and control groups.	[117]
MDD: anxiety	MDD genetic factors overlap with anxiety	[118]
MDD or BD: white matter integrity	MDD with white matter integrity	[126]
MDD: brain structure	MDD and reduced cortical thickness of the amygdala-medial prefrontal cortex	[127]
Extraversion or neuroticism: anxiety, MDD or psychological distress	Mixed results with nominal significance reflecting sample size	[60]
5 personality traits: MDD or BD	Neuroticism with MDD. Extraversion BD	[61]
Schizophrenia: BD	Schizophrenia with BD	[8]
5 Psychiatric disorders (schizophrenia, BD, MDD, ADHD and ASD each uses as discovery and target	Schizophrenia with BD, MDD and ASD BD with MDD and ASD. (All reciprocal)	[53,58]
Behavioral disinhibition, alcohol use, drug use, nicotine use/dependence	Shared genetic factors	[125]
Schizophrenia and/or BD: ADHD	GPRS from schizophrenia associated with ADHD but stronger association when discovery was schizophrenia and bipolar disorder	[59]

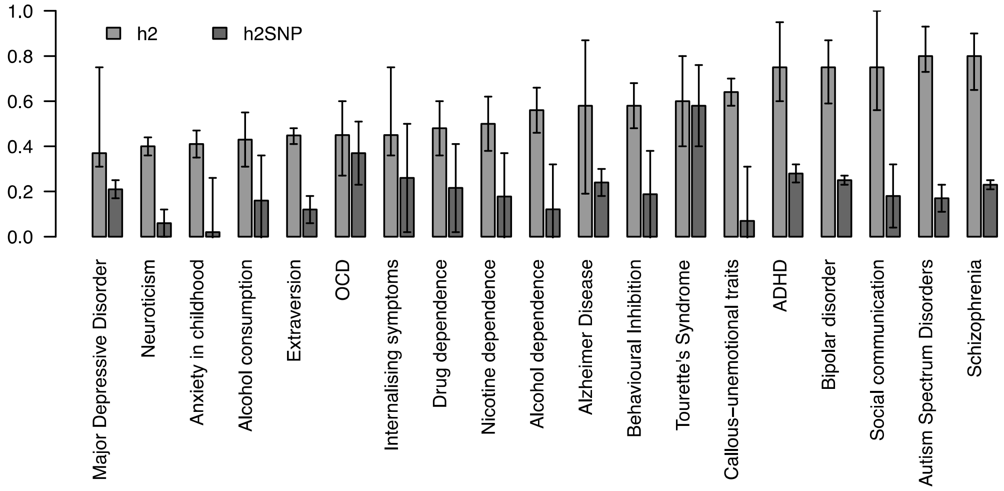
BD: Bipolar disorder. MDD: Major depressive disorder. ASD: Autism spectrum disorder. See Table S3 for more details.

Table 3: Overlap in genetic risk factors between disorders/traits and subtypes, GPRS studies

Discovery: Target Phenotypes	**Evidence for overlap**	**Reference**
ADHD: ADHD with/without conduct disorder	Polygenic risk scores are higher in those with conduct disorder	[63]
ADHD: ADHD-related traits in a population sample	ADHD explained significant variance in the traits measured in a population sample implying that ADHD is represents the extreme end of traits present in the general population	[64]
Schizophrenia: schizophrenia spectrum including unaffected relatives	Significant prediction with the most significant result for the narrow phenotype and the less significant result for being an unaffected relative	[128]
3 schizophrenia symptom dimensions: schizophrenia	Negative/disorganized dimension is most associated with schizophrenia	[129]
Schizophrenia: bipolar subtypes	Schizophrenia derived GPRS discriminate between schizoaffective bipolar disorder and non schizoaffective bipolar disorder but not between bipolar disorder with and without psychosis	[130]
Schizophrenia: positive, negative, disorganization, mania and depression symptom dimensions.	Schizophrenia GPRS are associated with all symptom dimensions in case vs control analysis, but not for symptom dimensions within case or control sample separately.	[120]
Schizophrenia: schizophrenia + schizoaffective disorder + psychotic bipolar disorder	Schizophrenia GPRS are associated with a broad psychosis phenotype	[128]

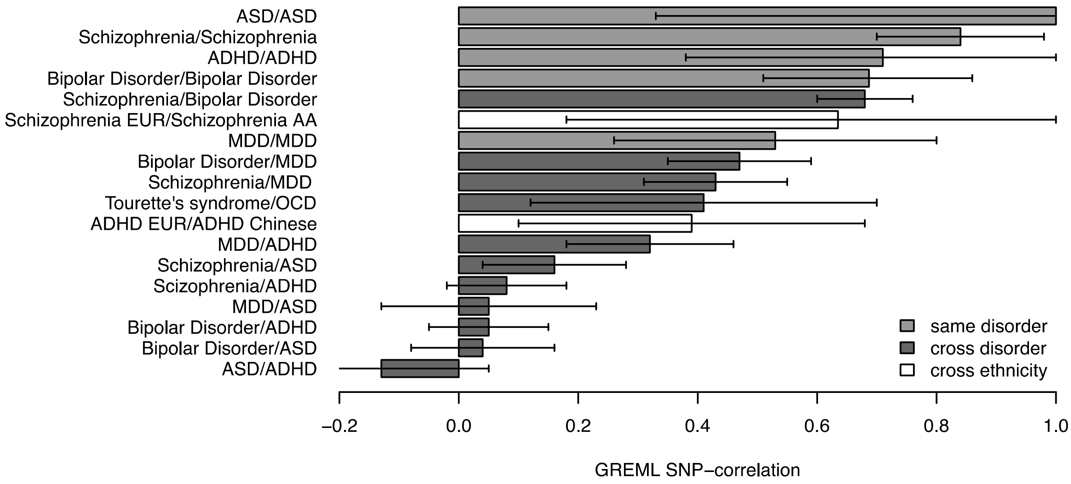
See Table S4 for more details

Figure 1. Heritability of liability from family studies and GREML SNP-heritability for psychiatric disorders and related traits.



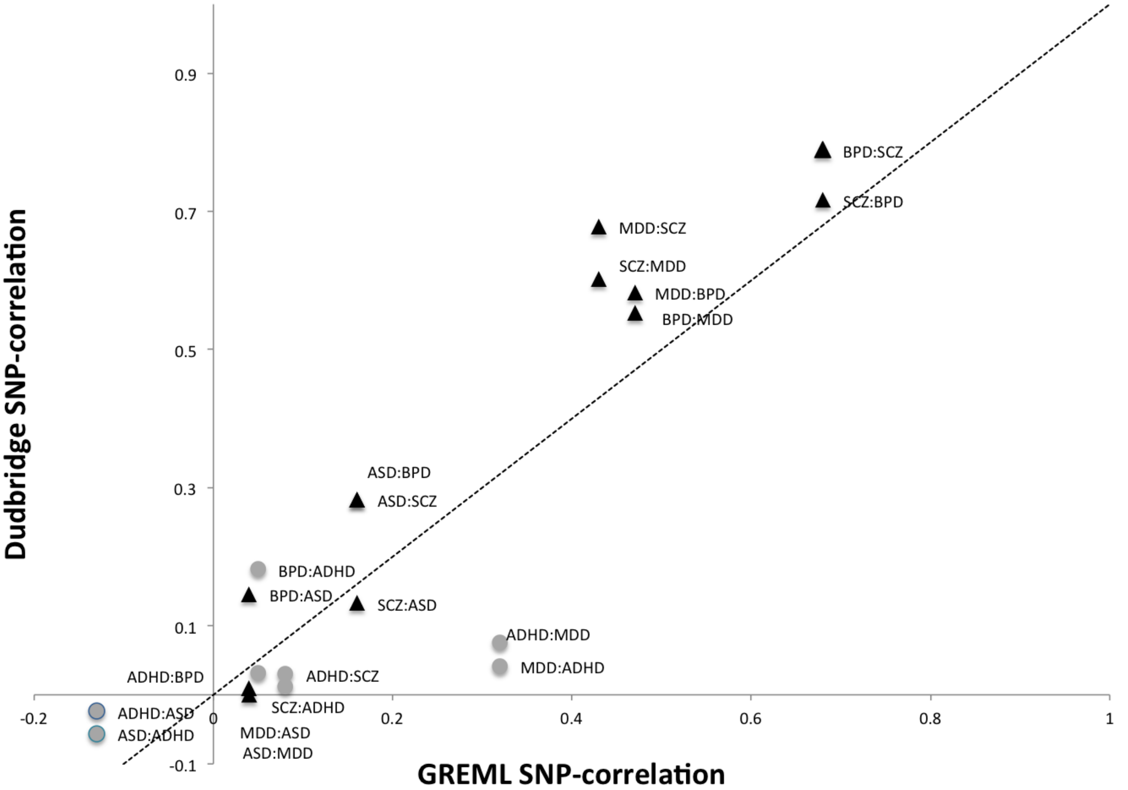
For more details and references see Table S1. For heritabilities the bars show a mixture of 95% confidence intervals from meta-analysis and of reported ranges. For SNP-heritabilities the 95%CI are approximated as the estimate ± 1.96 s.e. OCD: Obsessive compulsive disorder. ADHD: Attention deficit hyperactivity disorder.

Figure 2. Quantifying the genetic relationship between independent data sets through the SNP-correlation[41,44,51,52].



For 95%CI are approximated as the estimate ± 1.96 s.e. See Table S5 for more details.

Figure 3. Relationship between SNP correlation estimated from GREML and published in[41] and SNP-correlation estimated from the Dudbridge method[34] using the GPRS results published in[53], discovery disorder: target disorder. The same data were used (black triangles) except for analyses using ADHD (grey circles) for which more data was used in[41]. Dotted line y=x. Correlation between GREML and Dubridge estimates = 0.88. See Table S6 for more details. The Dudbridge correlation estimates are calculated using univariate GREML estimates of SNP heritability.



ADHD: attention deficit hyperactivity disorder, ASD: autism spectrum disorder, BPD: bipolar disorder, MDD: Major Depressive Disorder, SCZ: Schizophrenia.