Title

PA-FGRS is a novel estimator of pedigree-based genetic liability that complements genotype-based inferences into the genetic architecture of major depressive disorder

Authors (Working)

Morten Dybdahl Krebs[1,2,+], Kajsa-Lotta Georgii Hellberg[1,2], Mischa Lundberg[1,2], Vivek Appadurai[1,2], Henrik Ohlsson[3], Emil Pedersen[4], Jette Steinbach[4], Jamie Matthews[5], Sonja LaBianca[1,2], Xabier Calle[1,2], Joeri J. Meijsen[1], iPSYCH Study Consortium[#], Andres Ingasson[1,2], Alfonso Buil[1,2], Bjarni J. Vilhjálmsson[2,4,6], Jonathan Flint[7], Silviu-Alin Bacanu[8,9], Na Cai[10], Andy Dahl[11], Noah Zaitlen[5,12], Thomas Werge[1,2], Kenneth S. Kendler[8,9], Andrew J. Schork[1,2,13,+]

Affiliations

1. Institute of Biological Psychiatry, Mental Health Center - Sct Hans, Copenhagen University Hospital - Mental Health Services CPH, Copenhagen, Denmark
2. The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen, Denmark
3. Center for Primary Health Care Research, Lund University, Malmö, Sweden
4. NCRR - National Centre for Register-Based Research, Business and Social Sciences, Aarhus University, Aarhus V, Denmark
5. Department of Computational Medicine, University of California, Los Angeles, California, USA
6. Department of Biomedicine, Aarhus University, Aarhus, Denmark
7. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA
8. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA
9. Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA
10. Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany
11. Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, USA
12. Department of Neurology, University of California, Los Angeles, California 90024, USA
13. Section for Geogenetics, GLOBE Institute, Faculty of Health and Medical Science , Copenhagen University

iPSYCH Study Consortium authors can be found in the supplementary tables.

+ Please address correspondence to:
Morten.Dybdahl.Krebs@regionh.dk
Andrew.Joseph.Schork@regionh.dk

Abstract

The state-of-the-field in complex disorder genetics is marked by large-scale, biobank-ascertained data sets that integrate broad demographic, health, survey, and genetic data. Leveraging the richness of this data can advance our understanding of complex disorders by improving predictions, describing etiology, and augmenting gene-mapping. These themes are especially, but not exclusively, relevant to Major Depressive Disorder (MDD), a leading cause of disability for which genetic predictors underperform, clinical heterogeneity is etiologically enigmatic, and missing heritability persists. The iPSYCH 2015 case-cohort presents a unique opportunity to integrate genotypes, register-based clinical outcomes, and extended genealogies of 30,949 MDD cases and 39,655 random population controls. To make full use of our data, we introduce the Pearson-Aitken framework for Family Genetic Risk Scores (PA-FGRS) to estimate individuals' liabilities for major depressive disorder from extended genealogies of partially observed relatives. PA-FGRS extends previous methods by accounting for censoring, leveraging distant relatives, and utilizing a flexible, model-based approach amenable to analysis and extension - advantages we highlight in simulations. Combining PA-FGRS with genotype data improves classification, replicates and extends known genetic contributions to clinical heterogeneity, and increases power for genome-wide association studies. This study combines novel analyses of unique data and can serve as a model for studies of other outcomes.

Introduction

The state-of-the-field for complex disorder genetics is defined by the emergence and incorporation of large-scale, biobank-ascertained data sets (e.g., BioBank Japan[1], deCODE genetics[2] , iPSYCH[3,4], UKBiobank[5], etc.). These resources combine broad demographic, health, survey, and various genetic data on hundreds of thousands of individuals. Studies in population-level biobank data can take advantage of rich variable spaces at large scale to, e.g., emphasize genetic heterogeneity of psychiatric outcomes[6,7], combine family history data and genotype data to power genome-wide association studies (GWAS)[8,9], and integrate polygenic risk scores (PGS) with other predictors[10,11]. In contrast, large meta-analytic genetic studies, despite their unquestionable impact, have a necessarily narrower scope. To achieve requisite inferential power, they often focus on few variables, ascertained inconsistently across cohorts, and with limited opportunities for contextual perspectives. As such, there is both an opportunity and need for studies and methods that integrate rich, complementary data to improve the power, interpretation, and context of genetic studies in complex disorders.

Major Depressive Disorder (MDD) can be viewed in many ways as a paradigmatic complex disorder. Twin and family studies establish genetic contributions to disorder liability, but estimates of heritability are more modest than for many other psychiatric disorders, ranging from ~0.3 to 0.5[12]. This is consistent with the notion that MDD may be both especially multifactorial and especially heterogeneous in its etiology and clinical presentation[13]. Despite a plethora of clinical descriptions, there remains a need to better understand the underlying etiology of clinical heterogeneity[14]. In addition, MDD is the most common mental disorder, with lifetime prevalence reported at ~0.2[12]. Together, these characteristics imply that GWAS of MDD are expected to be the least powered among psychiatric disorders[15,16]. Recent meta-analytic GWAS of MDD challenge this limitation with extreme sample sizes obtained by incorporating “light” definitions of MDD cases, i.e., a single item on a self-reported questionnaire[17,18]. This approach has uncovered hundreds of associated loci and increased the variance explained by results-derived polygenic scores (PGS)[17]. Nevertheless, it has been suggested that loci associated with “light” MDD may not elucidate core MDD biology[6] and, even with million-person MDD GWAS[6,18], PGS performance in MDD lags behind, e.g., Schizophrenia[19]. Current genetic studies would be complemented by deeper studies of single cohorts with cross-domain phenotyping that add context to the genetic architecture of MDD.

The iPSYCH 2015 case-cohort study[4] includes 30,949 cases of clinically defined MDD and 39,655 random population controls with imputed whole-genome genotyping. This sample is nested within the Danish national register system[20] and can be cross-linked with, e.g., registered measures of clinical

outcomes including severity, recurrence, age at first treatment, or comorbidity. Further, civil registers allow linkage of relatives through parent-offspring records that enable reconstructions of extended genealogies for all probands[21]. Broad register phenotyping is available for all probands and relatives during periods of legal residence in Denmark since 1969. Combining family history data with genotypes and clinical measures could complement current genetics studies of MDD by considering integrated predictions[11], genetic contributions to clinical heterogeneity[7,22,23], and powerful single cohort GWAS[8,9].

Currently, methods that transform patterns of binary diagnoses in a genealogy to continuous liability[24,25] scores in each family member can not accommodate the full extent of iPSYCH data. LT-FH[9] and LT-FH++[26] are related methods that use resampling to estimate posterior mean liabilities of relatives. They consider first degree relative records thereby excluding many iPSYCH relatives and potentially confounding estimates strongly with familial environment. These methods were developed and applied with the aim to improve GWAS. So et al[27] developed a method based on the Pearson-Aitken (PA) selection formula[28], a simple analytical procedure for calculating an expected liability score for a proband conditional on relative phenotypes in arbitrarily structured genealogies, assuming all relatives are fully observed. A more flexible, resampling-based extension of this model was proposed but is computationally prohibitive at scale[29]. These approaches were developed and applied with a focus on trait predictions. Kendler et al.[22] introduced family genetic risk scores (FGRS), kinship weighted sums of the disease status of relatives with corrections for familial environment, censoring, and other covariates. FGRS accommodates extended genealogies and censored records, but it is not based on a fully described model and FGRS does not fully account for kinship among relatives of probands. FGRS was developed and has been applied to describe genetic heterogeneity within and across disorders. Current methods estimating individual liability from genealogies would not take full advantage of iPSYCH data which includes extended genealogies of only partially observed relatives.

We introduce a new method, Pearson-Aitken Family Genetic Risk Scores (PA-FGRS), validate it under simulations, and apply it to the iPSYCH2015 case-cohort data to study major depressive disorder (MDD). We demonstrate that PA-FGRS improves inference along three lines of inquiry: 1) improving classification of MDD in the context of PGS, 2) highlighting robust genetic contributions to clinical heterogeneity in MDD, and 3) improving power in a large single cohort GWAS of MDD. Our applications confirm, add context to, and extend important recent methodological advances and their applications in similar data. The PA-FGRS framework is extensible, robust, easy to implement, and can be applied across biobank data resources or to pursue similar aims with other complex disorders.

Results

The iPSYCH 2015 MDD case-cohort genealogies are complex and contain a wealth of information

The iPSYCH 2015 case-cohort sample ascertained 141,265 probands from the population of individuals born in Denmark between May 1st 1981 and Dec 31st 2008 (N=1,657,449) by cross-linking the Danish civil register[30] (CPR) and neonatal biobank[31]. The CPR includes all individuals who have legally resided in Denmark since its establishment in 1968 and each proband is associated with parental identifiers, where known. We follow Athanasiadis et al.[21] in using mother-father-proband connections to reconstruct extended genealogies (Online Methods) of 141,265 iPSYCH2015 probands, identifying 2,066,657 unique relatives that span up to nine generations (birth years range: 1870s to 2016; Figure 1a, Supplementary Figure S1). Of the 20,071,410 relative pairs identified, 24,773 pairs were of iPSYCH probands genotyped on the same array. The pedigree inferred kinship was highly correlated with observed SNP-based kinship (r=0.969, Figure 1B), siblings sharing recorded one parent with the other missing tended to be half-siblings (Supplementary Figure S2), and we infer that approximately 45% of same-sex twins are monozygotic (Supplementary Figure S3). For the 141,265 probands, genealogies included 99.5% of parents, 82.0% of grandparents, and 7% of great-grandparents. As only relatives sharing a common ancestor alive and legally residing in Denmark after the establishment of the civil register can be connected, our ability to capture distant relatives is limited. The number of relatives identified per proband varied considerably (Figure 1C). Clinical diagnoses can be aggregated for all relatives during periods of legal residence within Denmark from 1968 with in-patient psychiatric contacts recorded from 1969 to 1994 using ICD-8 and ICD-10 from 1994 onwards, and since 1995 both in- and out-patient contacts recorded (Figure 1D,E). There is a wealth of high-quality psychiatric family history for each genotyped proband (Figure 1), but relatives are neither completely nor consistently observed.

PA-FGRS is a flexible, powerful framework for estimating individual liability scores

PA-FGRS estimates the expectation of a proband's genetic liability from an arbitrarily structured genealogy, assuming the outcome results from a thresholded latent Gaussian liability (Figure 2). As input PA-FGRS takes a kinship matrix, the diagnostic status and age (at censoring, diagnosis, or end of follow-up) for each pedigree member, a phenotype heritability, and lifetime sex by birth year-specific cumulative incidence. In a first step, each pedigree member is assigned an initial liability of 0 with variance 1. Then we consecutively condition on observations of other relatives, r_1 , …, r_n, updating all expected liabilities for each. This is done by first updating the expected liability of a selected relative, $r_{(i)}$,

estimating their expected liability given their prior liability distribution, disease status, age and the lifetime incidence estimate. Then we update the liabilities of all remaining relatives, r_{i+1}, …, r_n, according to the PA-selection formula[28] and a modified kinship matrix (Supplementary Figure S4). An optional final step updates the proband liability on their own diagnostic status and age. This results in a continuous liability score that summarizes the genetic family history in the proband's pedigree.

Other methods have approached this problem, but with certain limitations critical to our use case. Notably, prior implementations[27,32] of the Pearson-Aitken (PA) selection formula[28] assumed fully observed individuals (i.e., no censoring). We addressed this by modeling individuals as mixture of truncated Gaussians, with mixture proportions derived from individual morbid risks, in the computation of initial liability (Online Methods). FGRS[22] followed this concept, but PA-FGRS takes a more formal approach that improves the efficiency by incorporating kinship relationships among relatives as well as between relatives and proband. This results in a better calibrated score and an estimate of the conditional liability variance (Online Methods).

Simulations demonstrate the advantages of PA-FGRS

We simulated 6,750,000 four-generational pedigrees with an average of nine relatives per proband (range 0-18), generating phenotypes from a liability threshold model (Online Methods). We compared the performance of PA-FGRS to that of FGRS[22], PA[27], and LT-PA[9] and found that PA-FGRS consistently produced the highest correlations with true liability across a range of heritabilities and trait prevalences (Figure 3C,D) and the relative gains were largest when heritability, prevalence, and censoring were high. Crucially, PA-FGRS was the only method that was well-calibrated in the presence of censored data (Figure 3E, Supplementary Figure 5). We found that all methods produced individual liability scores that were highly correlated (Figure 3A,B, $r > 0.8$), suggesting that they target similar latent constructs. Methods incorporating more similar information were more highly concordant, e.g., extended relatives (Figure 3a,b, $r > 0.89$) or extended relatives and censoring (Figure 3A, $r > 0.95$). A fully specified Gibbs sampling-based approach[29] produced nearly identical estimates to PA-FGRS (r=0.999, Figure 3a,b), suggesting PA-FGRS behaves near optimally. Our implementation of this Gibbs sampling approach, however, is computationally intractable at scale. (Supplementary Figure 6)

One limitation of methods that consider only first degree relatives[9,26] is that estimated genetic liabilities may capture effects of familial environment. This may be desirable if the goal is to optimize prediction[11,27], only, but less so if the goal is to make etiological inferences[22]. We repeated our simulations including a common environment component of variance shared among first degree

relatives (Figure 3F, Supplementary Figure 7) - a typical quantitative genetics model[33]. In these scenarios PA-FGRS (and other approaches) produce estimates of genetic liability that are correlated with the environmental liability (Figure 3F). An advantage of leveraging extended genealogies is that we can omit close relatives as a sensitivity test for undue influence. In these scenarios, liabilities estimated after excluding first degree relatives remained good estimators of genetic liability and were uncorrelated with environmental liability (Figure 3F). The flexibility of PA-FGRS can add important context to estimated liabilities that may be especially important when interpreting, e.g., profiles of liability scores[22,23] or when large familial environment effects are a concern.

PA-FGRS requires external estimates of specific population parameters, namely, lifetime prevalence and heritability. Providing inaccurate estimates results in miscalibrated estimated liabilities, but had modest impact on the correlation between estimate and the true liability in our simulations (Supplementary Figure 8). Regardless, estimating these parameters with reasonable precision is straightforward.

PA-FGRS contribute to classification of MDD over and above PGS

Both family history and PGS are known to explain liability for MDD. We constructed PA-FGRS from diagnoses of relatives, masking the disorder status of probands, in the iPSYCH2012 MDD case-cohort and iPSYCH2015i MDD case-cohort (Supplementary Figure 9) and predicted outcomes together with PGS (Online Methods, Figure 4A,B). Both genetic instruments significantly classified MDD in both cohorts: iPSYCH2012 (AUC_{PGS}=0.588 (0.583-0.594), $p= 3.7 \times 10^{-229}$; $AUC_{PA\text{-}FGRS}$= 0.598 (0.592-0.603), $p=4.9 \times 10^{-328}$) and iPSYCH2015i (AUC_{PGS}=0.573 (0.565-0.580), $p=7.8 \times 10^{-94}$; $AUC_{PA\text{-}FGRS}$=0.576 (0.569-0.583), $p=4.1 \times 10^{-136}$). Each genetic instrument contributed independent information, as demonstrated by their combined effects in a joint model being larger than marginal effects (iPSYCH2012: $AUC_{PGS+FGRS}$=0.630 (0.625-0.638) and iPSYCH2015i: $AUC_{PGS+FGRS}$=0.608 (0.601-0.615)).

Including PGSs for four other psychiatric disorders, schizophrenia (SCZ), bipolar disorder (BPD), autism spectrum disorder (ASD), and attention deficit/hyperactivity disorder (ADHD), improved the classification of MDD relative to models with MDD PGS only (iPSYCH2012: $AUC_{5\text{-}PGS}$=0.599 (0.594-0.604); iPSYCH2015i: $AUC_{5\text{-}PGS}$= 0.589 (0.582-0.596); Figure 4C,D). Similarly, incorporating PA-FGRS for the four other psychiatric disorders improved the classification of MDD relative to models with MDD PA-FGRS only (iPSYCH2012: $AUC_{5\text{-}PA\text{-}FGRS}$= 0.620 (0.614-0.625); iPSYCH2015i: $AUC_{5\text{-}PA\text{-}FGRS}$= 0.596 (0.589-0.603), Figure 4E,F). Combining all 10 predictors resulted in the best classification accuracy (iPSYCH2012:

$AUC_{5\text{-}PGS+5\text{-}PA\text{-}FGRS}$=0.648 (0.643-0.653); iPSYCH2015i: $AUC_{5\text{-}PGS+5\text{-}PA\text{-}FGRS}$= 0.626 (0.619-0.632), Figure 4G,H). These results demonstrate that combining genetic instruments that leverage different sources of genetic information improves classification of MDD.

Composite genetic profiles identify signatures of genetic heterogeneity in MDD

Individuals diagnosed with MDD demonstrate extensive clinical heterogeneity that may reflect etiologic heterogeneity. We used multinomial logistic regression to associate differences in clinical presentations of individuals diagnosed with MDD to their psychiatric genetic risk profiles (Online Methods, Figure 5). To leverage the complementarity of PGS and PA-FGRS, we defined composite genetic risk estimates for each disorder (e.g. BPD-score= β_{PGS}*PGS_{BPD}+ $\beta_{PA\text{-}FGRS}$*$PA\text{-}FGRS_{BPD}$, where β_{PGS} and $\beta_{PA\text{-}FGRS}$ are the estimated effect of the PGS and PA-FGRS on their natural outcome in a binomial logistic regression). Each composite psychiatric risk score was significantly larger in individuals diagnosed with MDD than in controls, across all subgroups (Figure 5; p<0.05). The estimated liabilities for bipolar disorder (BPD), schizophrenia (SCZ), autism spectrum disorders (ASD) and attention deficit/hyperactivity disorder (ADHD) tended to have smaller effects on MDD subgroups than on their natural outcome (i.e., β_{MLR}/β_{LR} < 1; the colored bars below dashed line in Figure 5; Online Methods), with the exception of BPD liability on conversion to a BPD diagnosis (β_{MLR}/β_{LR} =0.97 (0.90-1.04), Figure 5A).

Among 30,949 individuals diagnosed with MDD, those also diagnosed with BPD (N=1,477) had significantly (p < $1.4x10^{-3}$, adjusting for 35 tests) higher genetic liability for MDD ($p = 1.1 \times 10^{-12}$), BPD ($p = 4.7 \times 10^{-66}$), and SCZ ($p = 2.5 \times 10^{-6}$; Figure 5A). Among the 29,472 individuals diagnosed with MDD, but not BPD, the 7,205 also diagnosed with an anxiety disorder had higher genetic liability to MDD ($p = 4.9 \times 10^{-6}$) and SCZ ($p = 3.5 \times 10^{-12}$; Figure 5B). Individuals with recurrent depression (N=9,903) had higher liability to MDD ($p = 3.2 \times 10^{-12}$; Figure 5C) than those with single episode depression (N=19,569). Individuals treated for MDD in-patient ($N_{Hospitalized}$=5,815) had higher liability to MDD ($p = 6.2 \times 10^{-5}$) and BPD ($p = 8.1 \times 10^{-4}$) than those treated out-patient ($N_{Out\text{-}patient}$=12,432, Figure 5D). We did not observe any significant differences ($p > 1.4 \times 10^{-3}$) in the genetic liability score profiles of males vs females (N_{Female}=19,906, N_{Male}=9566; Figure 5E), based on age-at-first-diagnosis (Figure 5F), or based on diagnostic codes for severity (mild N_{Mild}=3,004, $N_{Moderate}$ =8,742, N_{Severe}= 2,391, $N_{Psychotic}$=856; Figure 5G).

Analyses were repeated for each genetic instrument, separately, (i.e., PGS or PA-FGRS only; Supplementary Figures S10-S11). The PGS-only and PA-FGRS-only results were highly similar to each

other (Pearson correlation: 0.95 (0.93-0.97); Figure 5H). Individual scores were less powerful than composite scores, however, (PA-FGRS-only mean $\log_{10}(p)$= 2.90; PGS-only mean $\log_{10}(p)$= 2.47; composite mean $\log_{10}(p)$= 4.24). Together this suggests that PGS and PA-FGRS may capture similar constructs and by combining the two we can increase power to detect genetic heterogeneity. Finally, to test for potential impacts of the familial environment on these inferences, we constructed PA-FGRS excluding nuclear family members (i.e., parents, siblings, half-siblings, and children). The overall trends were highly consistent with the full analysis (Figure 5I), albeit with reduced significance (Supplementary Figure 12). Genetic liability score profiles are associated with differences in clinical presentation of individuals diagnosed with MDD, often receive contributions from non-MDD liability scores, show parallel trends when considering PGS or PA-FGRS alone, and do not seem strongly influenced by familial environment.

GWAS on PA-FGRS liability values adds power to single cohort MDD GWAS

The multifactorial etiology and protracted age at onset of MDD imply that neither every case, nor every partially observed control will carry the same genetic risk and so studying genetic liability directly would boost power in GWAS (Supplementary Figure 13). We performed meta-analytic GWAS across the two iPSYCH MDD case-cohorts, 2012 (N_{cases}=17,518, N_{ctrl}=23,341) and 2015i (N_{cases}=8,323, N_{ctrl}=15,204, Supplementary Figure 9). We compare logistic regression on binary diagnoses to linear regression on PA-FGRS (Online Methods, Figure 6). GWAS for MDD PA-FGRS identified 3 independent loci (Figure 6A; index SNPs: rs16827974,β=0.014, p=2.9×10^{-8}; rs1040574, β=-0.011. p=3.3×10^{-8}; rs112585366, β=0.026, p=4.4×10^{-8}). These three variants and 24 of the 29 suggestive loci (false discovery rate <0.05) showed consistent sign in an independent MDD GWAS by Howard el al.[17] (excluding iPSYCH, Supplementary Table S2-S3). GWAS for MDD case-control status identified only first two of these loci as significant (Figure 6B; index SNPs: rs6780942, 8.5Kb from rs16827974 Beta=0.085, p=7.1×10^{-9} ; rs3777421 36.3Kb from rs1040574, β=-0.073, p=4.6×10^{-8}). These two variants and 24 of the 35 suggestive loci (false discovery rate <0.05) showed consistent sign in Howard el al.[17] (excluding iPSYCH, Supplementary Table S2-S3). The 28 independent, genome-wide significant index SNPs reported in Howard el al.[17] (excluding iPSYCH) have slightly, but significantly, larger test statistics in the GWAS on PA-FGRS (PA-FGRS mean χ^2= 4.55; case-control mean χ^2= 3.80; paired t-test p=0.018; Figure 6C).

To test for improved power to detect polygenes, we trained polygenic scores in each subcohort (iPSYCH2012 or iPSYCH2015i) using GWAS performed in the other cohort (iPSYCH2015i or iPSYCH2012). In both evaluation cohorts, PGS trained with PA-FGRS GWAS were modestly, but significantly better at

classifying MDD diagnosed individuals versus controls (2012: $AUC_{case\text{-}control\ PGS}$ = 0.537 (0.531-0.542), $AUC_{PA\text{-}FGRS\ PGS}$ =0.544 (0.538-0.550), test of differences: p=3.9×10^{-5}; 2015i: $AUC_{case\text{-}control\ PGS}$ = 0.556 (0.548-0.563), $AUC_{PA\text{-}FGRS\ PGS}$=0.548 (0.540-0.556), test of differences: p=2.1×10^{-7}; Figure 6D). Observed scale SNP-h^2 was slightly, but not significantly larger in the PA-FGRS GWAS ($h^2_{obs,PA\text{-}FGRS}$-$h^2_{obs,PA\text{-}case/ctrl}$=0.015 (-0.013-0.043); Figure 6E). GWAS on PA-FGRS and case-control had similar genetic correlations with external studies of MDD and other psychiatric disorders (Figure 6F). GWAS of PA-FGRS modestly improves power relative to GWAS of case-control status for detecting disease associated loci and polygenes.

Discussion

The emergence of large, biobank cohorts enables studies of complex disorders that can combine multiple data sources to provide extended context for the genetic architecture of complex disorders, such as MDD. In this study, we have developed a new method for estimating genetic liability scores from extended pedigree data where individuals may be only partially observed. Our PA-FGRS outperforms existing methods, especially in scenarios most relevant for the iPSYCH2015 case-cohort study. The liabilities we estimate complement genotype-based inferences into MDD in three parallel and important lines of inquiry: 1) classification, we show that PA-FGRS liabilities improve the classification of MDD when combined with state-of-the-field PGS, 2) descriptions of etiology, we show that genetic score profiles integrating PGS and PA-FGRS liabilities can identify genetic contributions to clinical heterogeneity in MDD associated with comorbidity, recurrence, and severity, 3) gene mapping, we show that GWAS performed on PA-FGRS scores have more power than GWAS on case-control status. Our method is highly flexible, easy to use, and could be applied across multiple other datasets and to ask similar questions of other complex traits and diseases. Here, we took a data-resource-first approach of describing the unique characteristics of a powerful resource and tailoring a novel method to fully accommodate its peculiarities, rather than discarding or censoring to accommodate existing approaches. This could reflect a complementary approach to lowest common denominator cross-cohort meta-analysis, especially as newer, larger, deeper, and necessarily more peculiar, data sets emerge.

The iPSYCH case-cohort study is special in the depth of phenotyping and pedigree data available for genotyped probands. PA-FGRS can take full advantage of the extended genealogies with a robust, flexible, easy to implement approach for computing individual liability scores from patterns of binary diagnoses in a genealogy. PA-FGRS incorporates relatives of greater distance than recent methods[9,26] and

handles censoring by treating partially observed controls as a morbid risk weighted mixture of a case and control. By formalizing the FGRS of Kendler et al[22] within PA-selection theory, we are able to gain efficiency and improve the calibration and interpretability of estimated liabilities. The straightforward mathematical formulations of PA-FGRS are amenable to analysis and extension, representing a major advantage and future potential. For example, we show how we can include and exclude the proband or close relatives to change the meaning and use cases of the liability score. The framework itself, however, could be extended to include covariance among relatives beyond additive kinship, covariance among multiple traits, or different thresholds for different phenotypes believed to result from different levels of the same underlying liability.

Genetic instruments, such as polygenic risk scores, have garnered enthusiasm as potentially useful clinical instruments, but currently, for most complex disorders, they do not predict well enough to be impactful. Just as when GWAS came to appreciate that a few variants in isolation were insufficient to describe the etiology of a complex disorder[34], one genetic risk score in isolation is unlikely to be sufficiently powerful to find impact in the clinic[35]. Here we show combining our family based liabilities and genotype-based PGS, from multiple disorders, can improve classification accuracy substantially. In cancer[36] or coronary artery disease[37], risk models incorporate multiple measures - health states, health traits, family history, and PGS. In psychiatry, this has been pursued in more limited contexts, (e.g.,[38]). Previous studies have found that combining parental history information and PGS improves the prediction accuracy[38], however, these studies only considered risk associated with parental MDD and did not leverage diagnoses in other relatives. Integrative models that combine multiple sources of genetic information, such as family history, estimated liability, and PGS along with exposure data have the potential to advance the clinical utility of risk assessment in psychiatry but will require large population data and integrative models.

It is common clinical knowledge that individual patients present with unique trajectories of symptoms and outcomes. Some previous studies have used PGS in iPSYCH data to look at associations with age of onset, severity, hospitalization[39] and recurrence,[40] but conclude that polygenic liability contributes minimally to heterogeneity in MDD. We observe several significant effects of polygenic liability on clinical heterogeneity, but with different genetic instruments. Our models calibrate estimated effect sizes differently, to better accommodate that the scale of noisy instruments, such as PGS, can misrepresent the effect of an underlying liability construct. From our results, we propose that genetic liability for MDD and BPD are both important and substantial contributors to the course of illness and treatment setting, and the prior pessimism could be a function of limits of the implemented genetic instruments. Similarly, we believe our effect calibration is the reason that we observe BPD genetic

liability to be significantly more important than SCZ liability for conversion from MDD to BPD, whereas a previous study[41] found similar magnitudes of effect for SCZ and BPD PGS. Our study, with more powerful genetic instruments and a unique effect size calibration, extends and clarifies the results of previous studies in similar data.

Our results also support and are supported by extensive work in the Swedish registers. Kendler et al.[22,23,42] used family genetic risk scores (FGRS) to show differences in genetic liability adjusted for family environment of siblings and parents is associated with progression to BPD[22], comorbid anxiety[23], recurrence[22], treatment setting[42], and age-at-onset[22] of MDD. Here, in independent yet comparable data, we replicate many of their findings. We confirm higher genetic liability to MDD among cases with recurrent depression using composite, PGSs-only, and PA-FGRS-only liability scores. We also replicate a higher liability to MDD and SCZ among MDD cases with comorbid anxiety using our composite and PGS-only scores. We also saw the same trend of higher liability to both MDD and BPD among hospitalized cases, here, using our composite and PGS-only scores. Kendler et al[42] showed higher BPD liability in male MDD cases, which was nominally significant in our study. We did not replicate results for age at onset, as we find no significant differences in our data, however, iPSYCH has a much reduced range for age of onset (iPSYCH: 15 to 35, Kendler et al: <22 to >69) and only includes secondary care treated (i.e., more severe) MDD. Our study supports and is supported by studies of independent, comparable Swedish register data, replicating evidence for genetic heterogeneity in MDD using an alternative model-based approach for family liability and by incorporating molecular PGS the genetic instruments.

The focus of many previous methods for estimating family based liability has been to improve the power of GWAS[9,26], but reported gains have typically been both significant and very modest. Consistent with simulations, the relative increase in power in observed in highly ascertained case-control data is smaller than what has been reported for population-based studies. This is likely because in population studies, especially for rarer disorders, most of the variance in liability is hidden *within controls*, whereas for highly ascertained data, most of the variance in liability remains *between cases and controls*. As such, little is gained by moving from binary to continuous measures. Despite this, we do see an increase in GWAS power that previous attempts to leverage family history in MDD GWAS using iPSYCH did not[26]. We also extend this previous study by showing a small but significant gain in PGS performance trained on GWAS of PA-FGRS relative to case-control status. Although we observe small gains in power for GWAS, which is consistent with other studies, the most powerful applications of family liability estimators may lie in classification or descriptions of etiology.

Our study should be interpreted in light of a few important limitations. Certain modeling choices

could affect the reliability of PA-FGRS. First, pedigree size varies substantially among individuals which causes a regression towards the mean liability in individuals with fewer relatives. Second, modeling the censoring process requires external information about age-of-onset curves for disease of interest - as do the other methods modeling censoring - and these may change over time. While reliable age-of-onset curves are available for the present register coverage, estimating age of onset curves for past decades, with different diagnostic systems and different register coverage is challenging. Third, our proposed model assumes that the true liability of cases with different age and calendar year of onset is the same. Others have proposed a model where the liability threshold varies according to the age and calendar year specific prevalence[26]. While we did not see clear associations between age-at-first-registration and MDD liability, optimizing modeling choices to impact the predictive accuracy and GWAS power (e.g. estimated heritability) is an interesting future direction. Fourth, our estimated liabilities included all diagnostic information available through Dec 31 2016 and so some relatives will have received their diagnosis later than the probands. Thus the estimated variance explained in liability to depression will not be representative of the variance explained by a PA-FGRS constructed from the information available at an earlier point in time, which is a point that should be kept in mind if the purpose of the predictor is to use it at first psychiatric contact or even earlier.

Here, we have taken a data-first approach to studying the genetic architecture of MDD by tailoring both our study aims and method development to the particular strengths and challenges of a unique data resource. Doing so resulted in a methodological increment with broad applicability and highlights the utility of integrating multiple sources of genetic data when considering trait predictions, etiological descriptions, and gene mapping.

References

1. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
2. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci Data* **4**, 170115 (2017).
3. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
4. Bybjerg-Grauholm, J. *et al.* The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *bioRxiv* (2020) doi:10.1101/2020.11.30.20237768.
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 (2020).
7. LaBianca, S. *et al.* Polygenic profiles define aspects of clinical heterogeneity in ADHD. *bioRxiv* (2021) doi:10.1101/2021.07.13.21260299.
8. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
9. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).
10. Kachuri, L. *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat. Commun.* **11**, 6084 (2020).
11. Hujoel, M. L. A., Loh, P.-R., Neale, B. M. & Price, A. L. Incorporating family history of disease

improves polygenic risk scores in diverse populations. *bioRxiv* 2021.04.15.439975 (2021) doi:10.1101/2021.04.15.439975.

12. Kendall, K. M. *et al.* The genetic basis of major depression. *Psychol. Med.* **51**, 2217–2230 (2021).
13. Flint, J. & Kendler, K. S. The genetics of major depression. *Neuron* **81**, 484–503 (2014).
14. Cai, N., Choi, K. W. & Fried, E. I. Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. *Hum. Mol. Genet.* **29**, R10–R18 (2020).
15. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34**, 254–257 (2010).
16. Wray, N. R. *et al.* Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol. Psychiatry* **17**, 36–48 (2012).
17. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
18. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
19. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
20. Thygesen, L. C., Daasnes, C., Thaulow, I. & Brønnum-Hansen, H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. *Scand. J. Public Health* **39**, 12–16 (2011).
21. Athanasiadis, G. *et al.* A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
22. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. Family Genetic Risk Scores and the Genetic Architecture of Major Affective and Psychotic Disorders in a Swedish National Sample. *JAMA*

Psychiatry **78**, 735–743 (2021).

23. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. Impact of comorbidity on family genetic risk profiles for psychiatric and substance use disorders: a descriptive analysis. *Psychol. Med.* 1–10 (2021).
24. Wray, N. R. & Visscher, P. M. Quantitative genetics of disease traits. *J. Anim. Breed. Genet.* **132**, 198–203 (2015).
25. Wright, S. An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs. *Genetics* **19**, 506–536 (1934).
26. Pedersen, E. M. *et al.* Accounting for age of onset and family history improves power in genome-wide association studies. *Am. J. Hum. Genet.* **109**, 417–432 (2022).
27. So, H.-C., Kwan, J. S. H., 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**, 548–565 (2011).
28. Aitken, A. C. Note on selection from a multivariate normal population. *Proc. Edinb. Math. Soc.* **4**, 106–110 (1935).
29. Campbell, D. D., Li, Y. & Sham, P. C. Multifactorial disease risk calculator: Risk prediction for multifactorial disease pedigrees. *Genet. Epidemiol.* **42**, 130–133 (2018).
30. Pedersen, C. B. The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).
31. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
32. Mendell, N. R. & Elston, R. C. Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics* **30**, 41–57 (1974).
33. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer, 1998).
34. 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).

35. Schork, A. J., Schork, M. A. & Schork, N. J. Genetic risks and clinical rewards. *Nature genetics* vol. 50 1210–1211 (2018).
36. Lee, A. *et al.* BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).
37. Goff, D. C., Jr *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S49–73 (2014).
38. Agerbo, E. *et al.* Risk of Early-Onset Depression Associated With Polygenic Liability, Parental Psychiatric History, and Socioeconomic Status. *JAMA Psychiatry* **78**, 387–397 (2021).
39. Musliner, K. L. *et al.* Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population. *JAMA Psychiatry* **76**, 516–525 (2019).
40. Musliner, K. L. *et al.* Polygenic Liability and Recurrence of Depression in Patients With First-Onset Depression Treated in Hospital-Based Settings. *JAMA Psychiatry* **78**, 792–795 (2021).
41. Musliner, K. L. *et al.* Polygenic Risk and Progression to Bipolar or Psychotic Disorders Among Individuals Diagnosed With Unipolar Depression in Early Life. *Am. J. Psychiatry* **177**, 936–943 (2020).
42. Kendler, K. S., Ohlsson, H., Bacanu, S., Sundquist, J. & Sundquist, K. Differences in genetic risk score profiles for drug use disorder, major depression, and ADHD as a function of sex, age at onset, recurrence, mode of ascertainment, and treatment. *Psychol. Med.* **53**, 3448–3460 (2021).

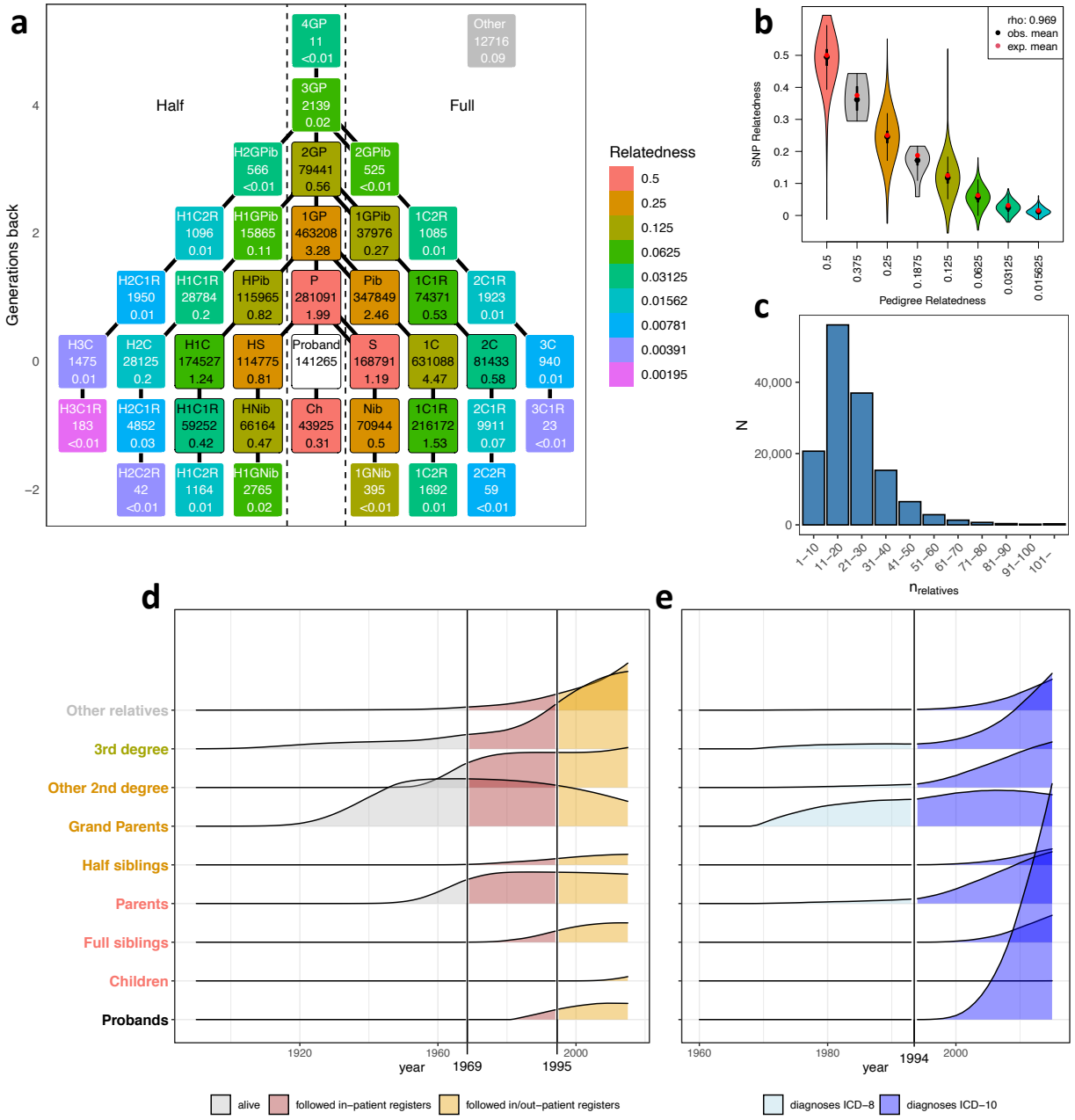

Figure 1 | The iPSYCH-2015 MDD case-cohort genealogies contain a wealth of information.
a) For each of the 141,265 probands in iPSYCH2015, there are a number of available relatives, here reported as a total across all and average per proband pedigree. **P,** parents; **S,** siblings; **Ch,** children; **1GP,** grandparents; **Pib**(lings), aunts and uncles; **Nib**(lings), nieces and nephews; **iCjR** i^{th} cousin, j^{th} removed; **H-,** half; **Other,** relative types not in the figure (including 702 double 1C, 2722 twin Pib, 5158 twin 1C and 1561 twin 1C1R.) **b**) SNP-based relatedness is highly correlated with that inferred from the genealogy. **c)** The number of relatives linked to each proband varies considerably. **d**) The amount of follow up, depicted as the proportion of total number of person-years lived in Denmark by all probands and their relatives, varies by relative type (y-axis), year of observation (x-axis), and register era (color). **e**) The cumulative proportion of individuals with a depression diagnosis stratified by relative type and colored by the register era.

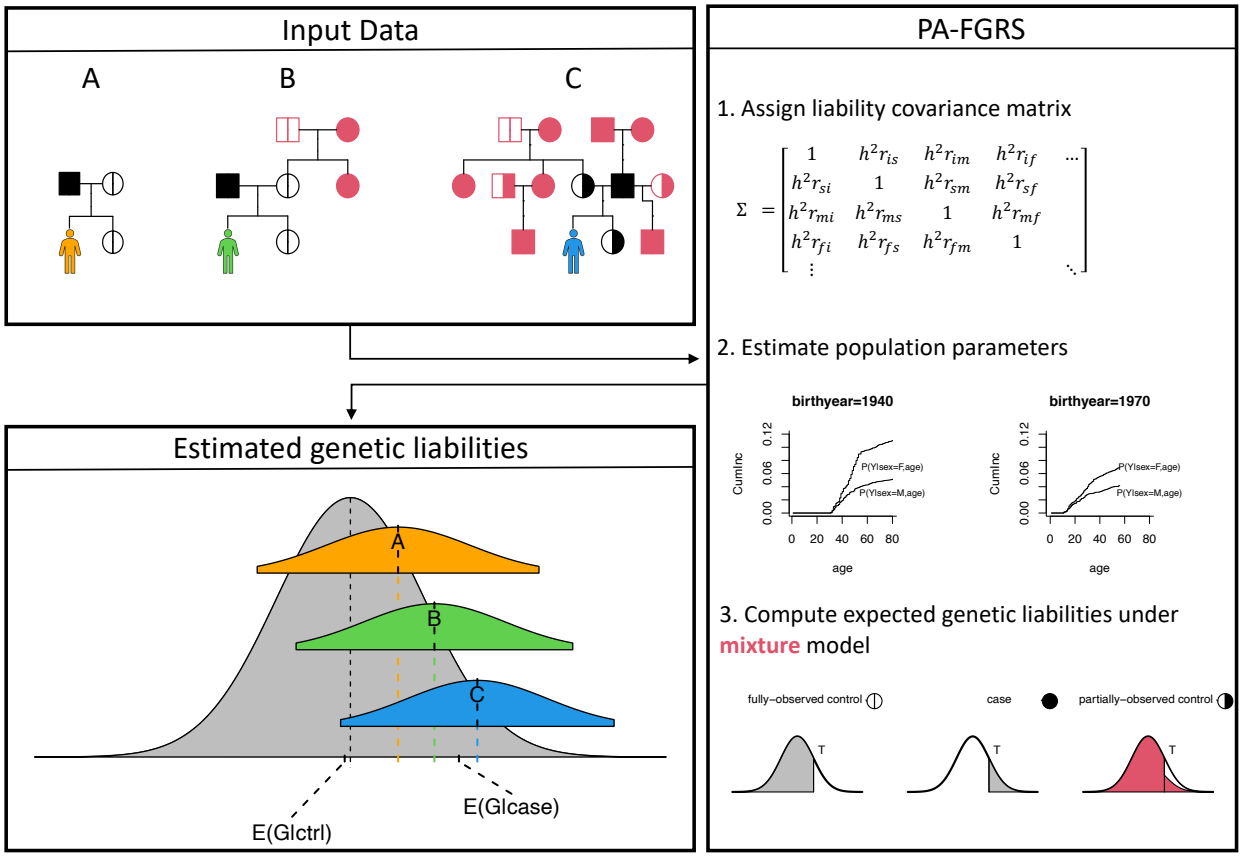

Figure 2 | PA-FGRS estimates liabilities from records of family history and disease specific population parameters.

PA-FGRS is a novel method for converting patterns of disease in arbitrarily structured pedigrees where relatives may only be partially observed into an estimate of latent disease liability for a proband. *Input data* for a proband can be a simple, fully observed pedigree (yellow proband), an extended, fully observed pedigrees (green proband), or an arbitrarily structured pedigree where many relatives are only partially observed (blue proband). *PA-FGRS* combines an assumed form for covariance in liabilities among relatives (1) with estimated, covariate stratified cumulative incidence curves (2) in a novel extension of the Pearson-Aitken selection formulas that models partially observed controls as a mixture of liability thresholded cases and controls. *Estimated genetic liabilities* are assigned to each proband and determined by the unique configuration of their pedigree. Proband liabilities (colored) are shown against a gray density that is the assumed population distribution of genetic liability where E(G|case) and E(G|ctrl) indicate the expected mean liability of a case and control, respectively.

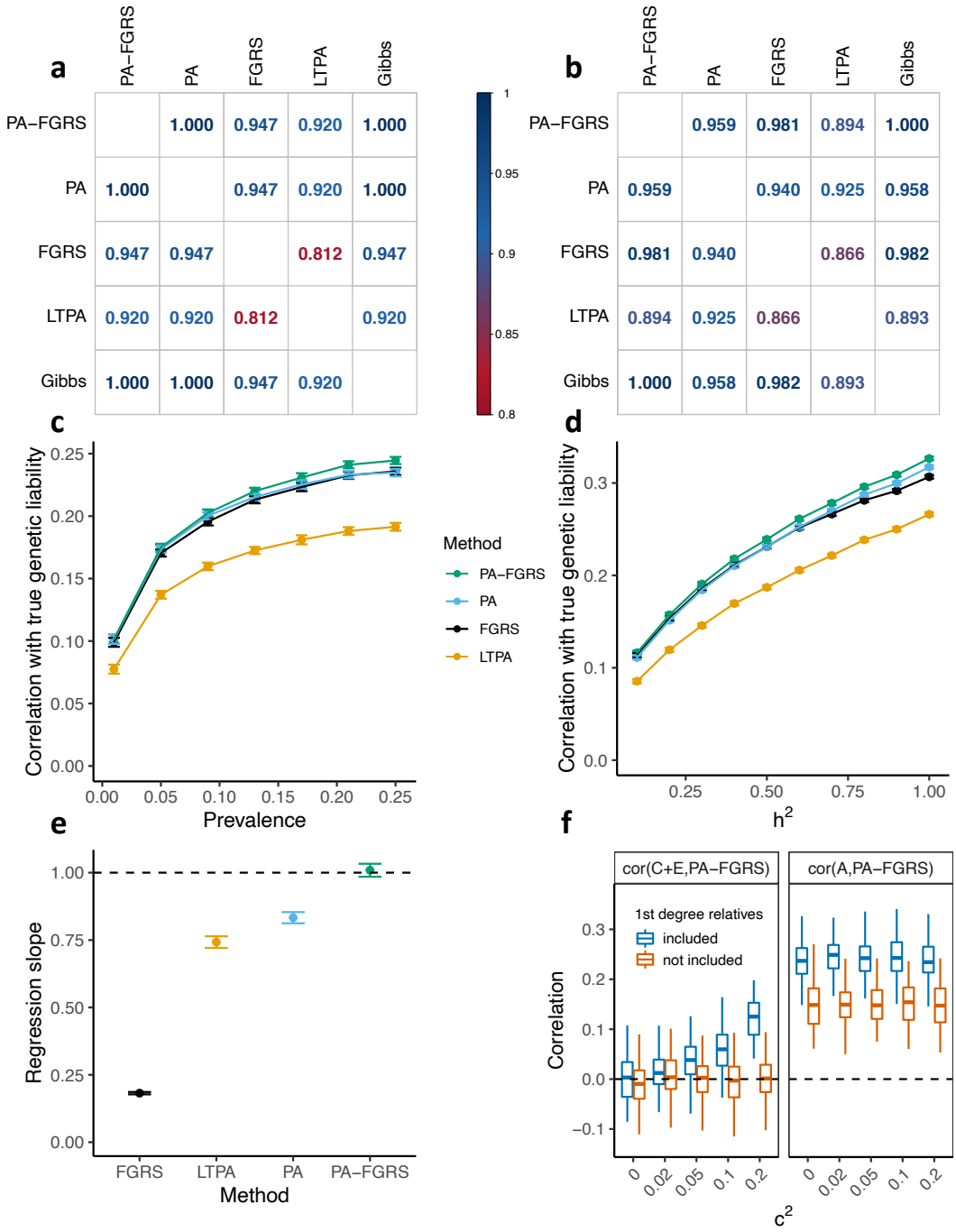

Figure 3 | PA-FGRS outperforms other methods in simulations.
PA-FGRS liabilities are correlated with those from other methods both (**a**) when all relatives are fully observed or **b**) when younger relatives are only partially observed (i.e., censored). In simulations, PA-FGRS shows the largest correlation with true genetic liability under **c**) varying trait prevalence and **d**) varying trait heritability. **e**) Linear regression of estimated liability on true liability shows PA-FGRS estimates are better calibrated estimates. **f**) Estimated liabilities from PA-FGRS are correlated with environmental components of variance when traits receive substantial contributions of family environment, but this can be diminished at the cost of reduced power (i.e. reduced correlation with genetic components) by removing confounded (i.e., nuclear) relationships. Panels c-e show mean and 95%-confidence interval across simulations, while f shows median, range and interquartile range across simulations.

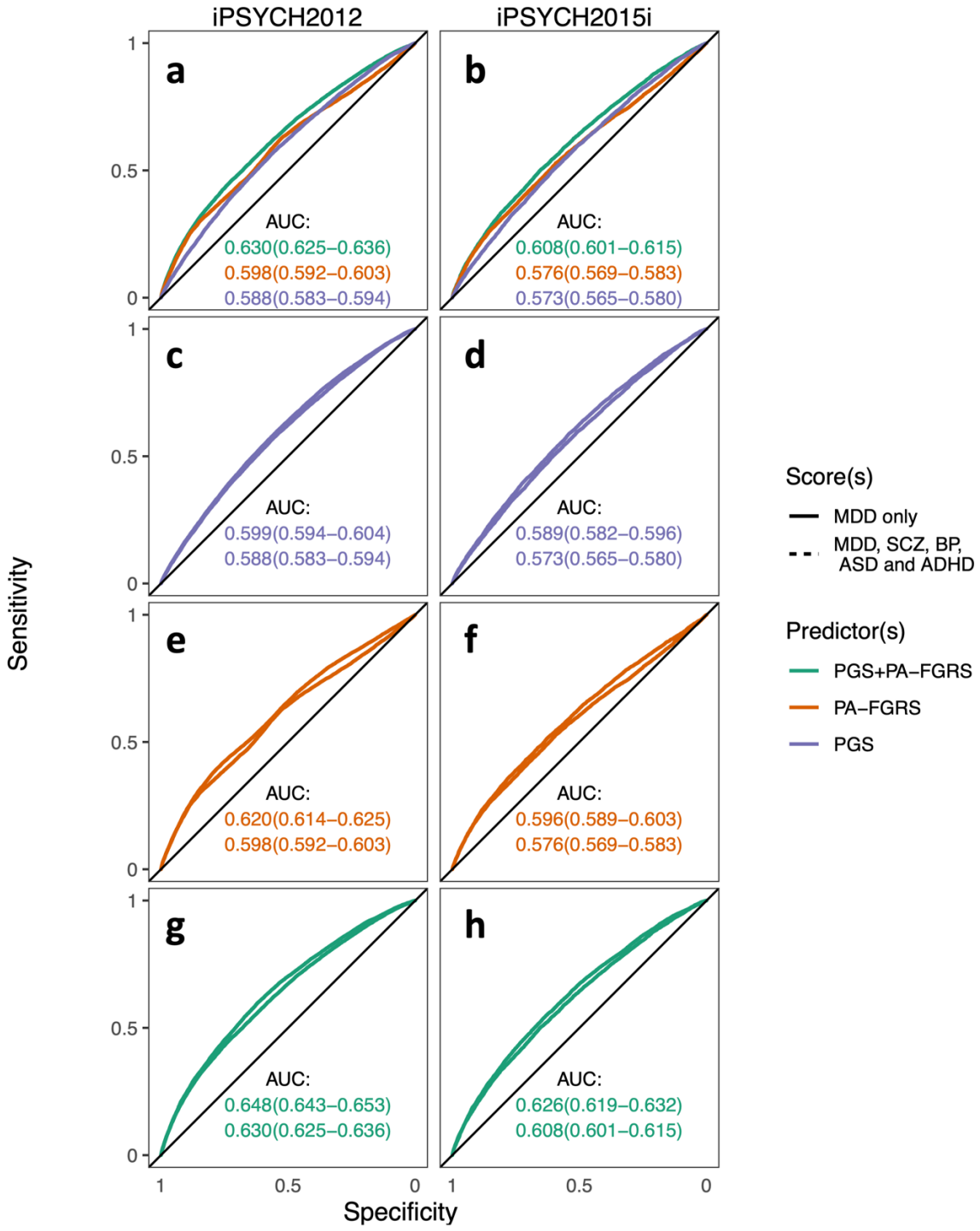

Figure 4 | PA-FGRS and PGS are complementary predictors of MDD
a,b Combining PA-PGRS-MDD and PGS-MDD improves prediction of MDD the iPSYCH-2012 (**a;** N_{cases}=20,632, N_{ctrl}=23,870) and iPSYCH-2015i (**b;** N_{cases}=10,317, N_{ctrl}=15,785) case-cohorts. **c,d** Using PGSs for five disorders improves prediction of MDD over only PGS-MDD in iPSYCH-2012 (**c**) and iPSYCH-2015i (**d**). **e,f** Using PA-FGRS for five disorders improves prediction of MDD over only PA-FGRS-MDD in iPSYCH-2012 (**e**) and iPSYCH-2015i (**f**). **g,h** Using PA-FGRS for five disorders and PGSs for five disorders improves prediction of MDD over only PA-FGRS-MDD and PGS-MDD in iPSYCH-2012 (**g**) and iPSYCH-2015i (**h**). **AUC** area under the receiver operating characteristic curve with 95%-confidence interval.

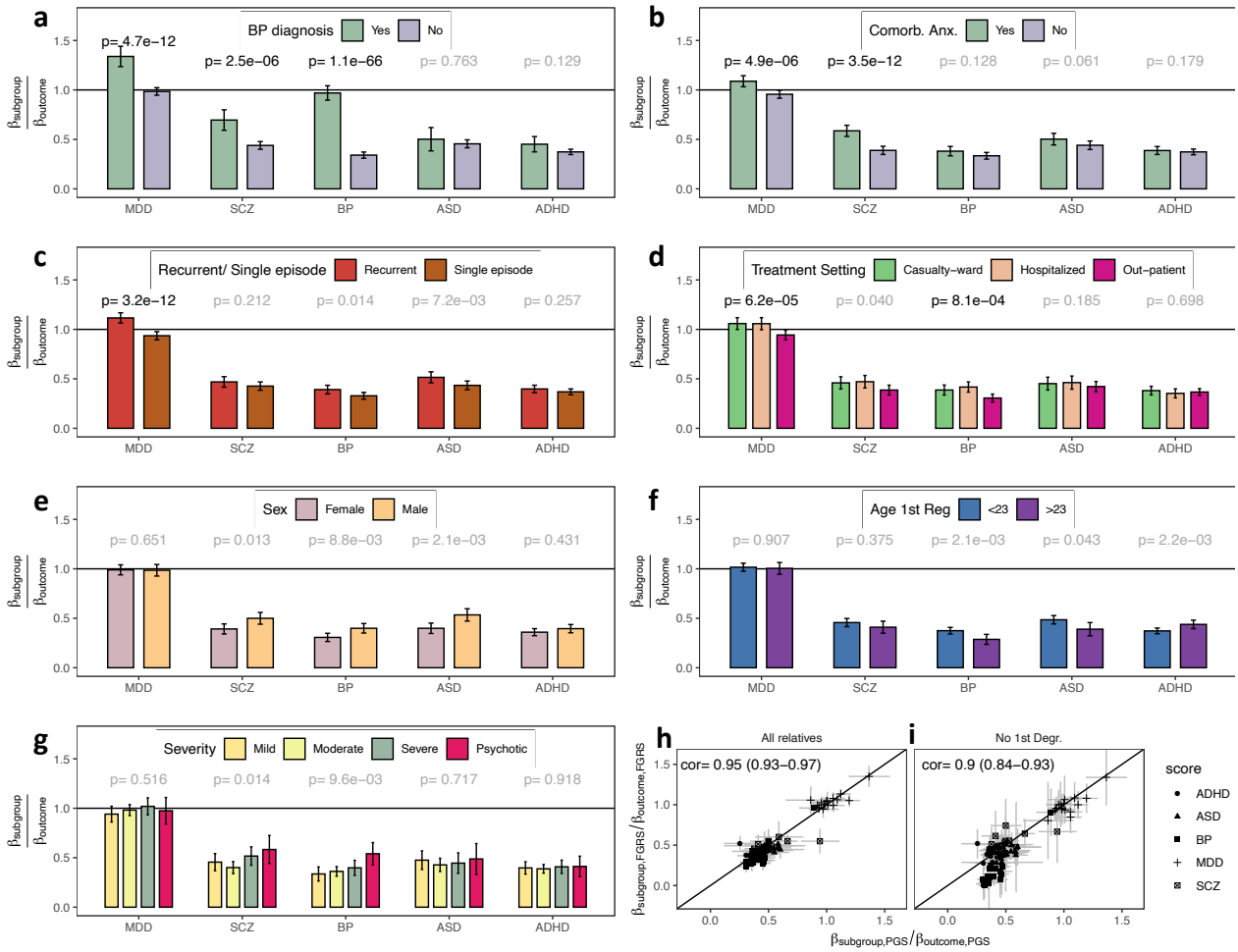

Figure 5 | Profiles of multivariable genetic liability are associated with clinical heterogeneity in MDD.
The figure shows the results of a multinomial logistic regression where the dependent variable is the categories of no-diagnosis and (**a**) MDD with/without bipolar disorder diagnosis, (**b**) MDD with/without comorbid anxiety, (**c**) single depressive episode and recurrent MDD (**d**) out-patient, casualty-ward and inpatient diagnoses of MDD (**e**) female MDD and male MDD, (**f**) first MDD diagnosis before/after age 23, (**g**), mild, moderate, severe, or psychotic depression. The independent variable is a composite estimate of genetic risk of one of five different mental disorders, constructed as a weighted sum of PA-FGRS and PGS. The y-axis indicates the estimated coefficient divided by the coefficient for the target diagnosis in a binomial logistic regression. **MDD** major depressive disorder, **SCZ** schizophrenia, **BPD** bipolar disorder, **ASD** autism spectrum disorders, **ADHD** attention-deficit/ hyperactivity disorder**, p** the probability of observing this data under the null-hypothesis (that all outcomes have the same coefficient). P-value in black indicates p<0.05/35. Beta-estimates and p-values are meta-analyzed across iPSYCH-2012 (N_{cases}≤20,632, N_{ctrl}≤23,870) and iPSYCH2015i (N_{cases}≤10,317, N_{ctrl}≤15,785) with the exception of panel **g** (Severity) which are from iPSYCH2012 only. Sample sizes for the individual analyses are provided in Table S3. Error-bars indicate 95% confidence intervals. PGS-only and PA-FGRS-only effects are highly consistent both, when using all relatives (**h**) and when excluding first degree relatives (**i**).

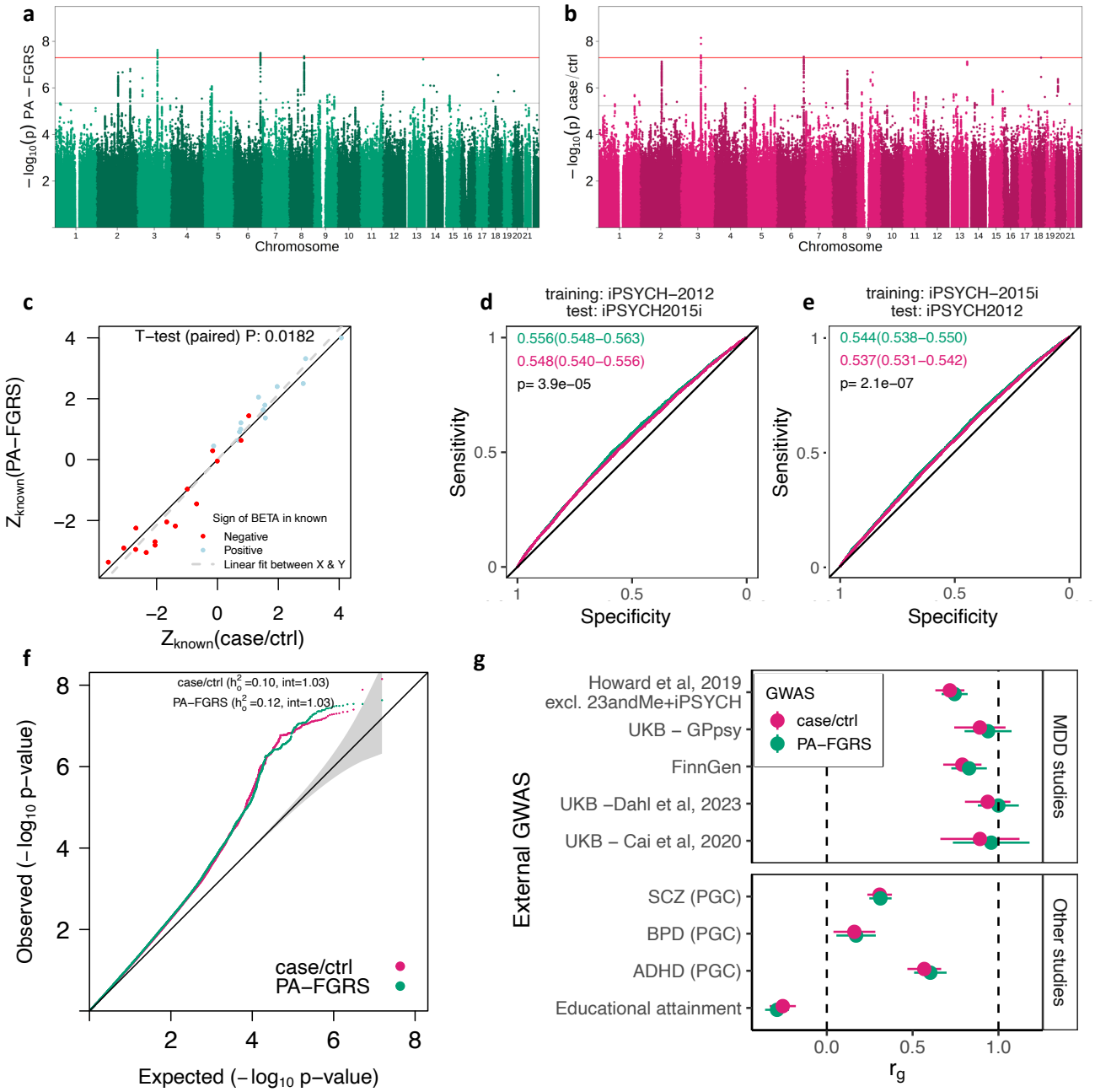

Figure 6 | PA-FGRS liabilities improve power for GWAS of MDD

a,b) Genome wide association studies (GWAS) of 25,841 cases and 38,545 controls using (**a**) PA-FGRS liability finds three independent genome-wide significant loci while (**b**) logistic regression (case/ctrl) finds two. **c)** PA-FGRS GWAS test statistics are more extreme (i.e., more significant) than case-control GWAS at index SNPs for 28 loci reported in a previous GWAS of MDD. **d,e)** PGS trained using PA-FGRS GWAS achieve higher classification accuracy that those trained on case-control GWAS in two independent evaluation cohorts. **f)** SNP-heritability estimated by LD-score regression analyses is slightly, but not significantly, larger for PA-FGRS GWAS while intercepts are equivalent. **g)** PA-FGRS and case-control GWAS show similar genetic correlations with external GWAS of MDD and related traits.