Polygenic prediction of the phenome, across ancestry, in emerging adulthood

Anna R. Docherty, Ph.D.*[1,2], Arden Moscati, Ph.D.*[2], Danielle Dick, Ph.D.[3-5], Jeanne E. Savage, Ph.D.[2], Jessica E. Salvatore, Ph.D.[2,3], Megan Cooke, Ph.D.[2], Fazil Aliev, Ph.D.[3,6], Ashlee A. Moore, M.A.[2], Alexis C. Edwards, Ph.D.[2], Brien P. Riley, Ph.D.[2], Daniel E. Adkins, Ph.D.[1,7], Roseann Peterson, Ph.D.[2], Bradley T. Webb, Ph.D.[2], Silviu A. Bacanu, Ph.D.[2], Kenneth S. Kendler, M.D.[2]

RUNNING HEAD: Genome-Phenome Prediction in Emerging Adulthood

1. Departments of Psychiatry & Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA
2. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University School of Medicine, Richmond, VA, USA
3. Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA
4. Department of Human & Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA
5. College Behavioral and Emotional Health Institute, Virginia Commonwealth University, Richmond, VA, USA
6. Department of Business, Karabuk University, Turkey
7. Department of Sociology, University of Utah, Salt Lake City, UT, USA

*Both authors contributed equally to the manuscript.

[+]Corresponding author: Departments of Psychiatry and Human Genetics, University of Utah School of Medicine, 501 Chipeta Way, Salt Lake City, Utah, USA. Tel. +1 763 516 7367, fax. +1 804 828 1471, email: anna.docherty@utah.edu

Article word count: 3,540 (excluding title page, abstract, references, figure captions, tables, author contributions, and financial disclosures)

Abstract

Background: Identifying genetic relationships between complex traits in emerging adulthood can provide useful etiological insights into risk for psychopathology. College-age individuals are under-represented in genomic analyses thus far, and the majority of work has focused on clinical disorder or cognitive abilities rather than normal-range behavioral outcomes.

Methods: This study examined a sample of emerging adults 18-22 years of age (N = 5,947) to construct an atlas of polygenic risk for 33 traits predicting real-world outcomes. Twenty-eight hypotheses were tested based on the previous literature on samples of European ancestry, and the availability of rich assessment data allowed for polygenic predictions across 55 psychological and medical phenotypes.

Results: Polygenic risk for schizophrenia in emerging adults predicted anxiety, depression, nicotine use, trauma, and family history of psychological disorders. Polygenic risk for neuroticism predicted anxiety, depression, phobia, panic, neuroticism, and risk for cardiovascular disease.

Conclusions: These results calcify unique links between genetic risk for schizophrenia, neuroticism, substance use, and important health factors in healthy early adulthood, and demonstrate cross-ancestry replication of these genetic relationships.

Keywords: phenome; genetic; polygenic; schizophrenia; neuroticism; cardiovascular

Introduction

Broad phenotyping can greatly enhance our understanding of the underlying structure of genetic vulnerability to psychiatric disorders. Recent genome-wide polygenic risk research has utilized a range of clinical traits (the "phenome") (Bulik-Sullivan *et al.* 2015; Krapohl *et al.* 2015; Hagenaars *et al.* 2016). This approach uses published summary statistics from large genome-wide association studies to calculate genome-wide polygenic scores (GPS) for many major disorders and clinically relevant traits. These scores are then used to predict a number of potentially informative psychiatric, psychological and physical health phenotypes. In conjunction with psychiatric genomic cross-disorder research (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013) examining the co-heritability of major psychiatric disorders (e.g., Docherty *et al.* 2016), this appears to be a promising method for elucidating a genetic framework for psychiatric disorders and discovering unexpected genotype-phenotype associations. However, previous research has not included GPS of eating, anxiety, and inflammatory disorders, personality, lipid levels and puberty traits in the prediction of outcome phenotypes (which have frequently focused on cognitive abilities) and has not examined samples between the ages of 17 and 35, the key age brackets of emerging and young adulthood.

This study applied such an approach to GPS (33 total) in a substantial genetic study of emerging adulthood outcomes. Emerging adulthood, a period starting at the age of 18 when adolescents begin to develop the roles and independence of adulthood, reflects a high-risk age range for the onset of many psychiatric and substance use disorders, including schizophrenia, affective disorders, anxiety disorders and alcohol and drug use disorders. Data from the National Comorbidity Survey Replication sample indicate that three quarters of all lifetime cases of DSM-IV diagnoses start by age 24 (Kessler *et al.* 2005), and WHO's World Mental Health data indicates that approximately three quarters of lifetime psychiatric disorders begin by the mid-20′s (Kessler *et al.* 2007).

The University Student Survey (called "Spit for Science", or S4S) was developed to identify risk factors for onset of mental health disorders with large-scale studies of genetic, environmental, and developmental influences. Discovery summary statistics from 33 genome-wide association studies (GWAS) were used to derive GPSs in this large sample of young adults (N = 5,947) across a range of psychiatric, psychological, and physical health traits (Table S1 in online supplementary materials). Expanding on previous research, twenty-eight hypotheses of genetic prediction were tested based on selected studies in past literature. Further, the availability of rich clinical assessment

data allowed for the calculation of polygenic predictions across a greater number of outcomes than has ever been studied previously, many of which were completely novel in phenomic studies. These included 55 psychiatric, psychological and medical phenotypes (listed in Table S2 in the online supplementary material).

Moreover, the GPS metrics were powerful enough to examine relationships across subsamples of different ancestries. While GWAS approaches require thousands of individuals to locate "hits," continuous polygenic scores require far smaller samples for adequate power. This sample was suitably diverse in ancestry to map the GPS-phenotype in young adults of European ancestry (EUR, N=3,016) and then to replicate these findings and GPS-GPS correlations across non-European ancestry groups including of African origin (AFR, N=1,339), native American origin (AMR, N=581), and East Asian (EAS, N=557), and South Asian origin (SAS, N=454). Separate association matrices were created for the empirically categorized AFR, AMR, EAS, and SAS samples and are provided here and in the supplemental figures available online.

We can learn a lot from the study of emerging adults over and above adolescent samples, as early behavioral patterns that may precede adult psychopathology can be studied, and new hypotheses about critical exposures and environmental risk factors can emerge. The results presented here reflect a polygenic modeling framework in a large young adult sample, and provides evidence that the integration of phenotypic and genotypic data will be useful in the prediction of negative health outcomes in emerging adults.

Methods

Sample Ascertainment and Phenotyping

Phenome-wide behavioral data (N=7,592) were drawn from young adults from the first three cohorts in S4S, samples drawn from a large urban university in the Mid-Atlantic United States, which included 5,947 unrelated individuals with genome-wide genotypes (Dick *et al.* 2014). The S4S sample does not overlap with any of the discovery GWAS samples used in these analyses. Details of participant ascertainment have been published elsewhere (Dick *et al.* 2014) but briefly, emerging adults ages 18-22 were recruited from a mid-sized university in the eastern United States, across multiple cohorts, for a campus-wide study of genetic and environmental factors contributing to alcohol and substance use. The protocol was approved by the university Institutional Review Board, and carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. Participants were 61.1% female with a mean age of 18.59 at first assessment. Representativeness of this sample is

strong and has been reported elsewhere (Dick *et al.* 2014). Assignment to ancestry group was empirically based on greatest similarity to 1000 Genomes Phase 3 super-populations. The present analyses included 55 traits from the domains of psychopathology, personality, health factors, and educational achievement (Table S2 in the online supplementary materials). All analyses included age, sex, and 10 ancestry principal components as covariates. Variables assessed at multiple occasions or in multiple cohorts were adjusted for number of assessments and cohort group. Sample sizes for each of the phenotypic measures are also provided in Table S2.

Genetic Risk Scoring

DNA collection, calling, and imputation is detailed elsewhere (Dick *et al.* 2014). We processed genotypes using standard quality control procedures followed by imputation of SNPs using the 1000 Genomes Project reference panel. After imputation and quality control, we included approximately 2.3 million variants into the polygenic scoring analyses. A GPS for each discovery phenotype was calculated using the summary statistics we obtained from 33 GWAS (Table S1 in the online supplementary materials). Python-based LDpred (Vilhjálmsson *et al.* 2015) was used for these analyses because of its ability to account for linkage disequilibrium (LD) structure (Krapohl *et al.* 2015) using our own large EUR test sample, and its use of all genetic variants (lack of specified p-value threshold for inclusion of the genetic variants in the GPS). LDpred allows for the modeling of LD based on LD in the discovery sample to weight the relative contributions of syntenic variants to the outcome phenotype. LDpred uses postulated proportions of causal variants in the genome as Bayesian prior probabilities for GPS calculations, and we tested a range of different priors (proportions of 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001), as well as the model of infinite variants of infinitesimally small effect (Fisher, 1919) to construct scores.

Phenotype Prediction

A flowchart depicting the GPS-phenome cross-ancestry prediction and GPS-GPS correlation procedure is presented in Figure 1. Regressions were run using R to compare full (GPS, ten ancestry principal components, age, sex, cohort, and number of measurements when applicable) and restricted models where GPS was removed. Prior to the global analyses, a set of *a priori* hypotheses, gathered from previous research, were tested (Table 1). We elected to forgo experimental binning (into quantiles, for example) in order to minimize the number of exploratory analyses beyond regressions of GPS on the phenotypes. Multiple testing was corrected for using a False Discovery Rate (FDR) of 5% (Benjamini & Hochberg, 1995) within each ancestry group using the p.adjust function in R; the FDR is appropriate for an analysis designed to evaluate the pattern of relationships between many constructs because it

treats each combination of discovery phenotype, outcome, and LDpred prior level as an independent test. It should be noted that this multiple testing correction was conservative, not accounting for previously established associations or correlations between multiple prior levels tested in the same discovery phenotypes, to filter out any potentially spurious results.

Cross-Disorder GPS Partial Correlations and GPS-GPS Replication Hypotheses Across Ancestry

In addition to testing the GPSs prediction of the phenotypes, GPSs were also examined for correlations with each other in all ethnicities. These provide different results than genetic correlation estimates, but are intended to demonstrate that GPS scores are not independent, and that variance attributable to a particular discovery phenotype may be partially shared with another. This sharing may be due to common genetic factors between phenotypes, possible sample overlap, and error variance. GPS correlations have been previously reported in EUR, but this analysis added phenotypes such as cardiovascular and triglyceride factors. Correlation coefficients, *p*-values, and *q*-values (after correcting the p-values for the FDR of 5%) were derived for GPS partial correlations using R and adjusting for the ancestry principal components. We chose to use partial correlations in order to standardize the weights across phenotype and provide more direct comparisons of statistics for plotting purposes. Based on the cross-disorder psychiatric genomics findings to date (Bulik-Sullivan *et al.* 2015), we hypothesized significant GPS associations between schizophrenia (SZ) and bipolar disorder (BP), SZ and autism (AUT), SZ and major depressive disorder (MDD), BP and MDD, and AUT and attention deficit hyperactivity disorder (ADHD) across each of the ancestry groups (see Table 2).

Results

Genetic Profile Score-Phenotype Prediction

A Priori Replication Analyses

We evaluated previous cross-phenotype predictions based on recent work—for example, that age at menarche had an inverse association with obesity/body mass index (Bulik-Sullivan *et al.* 2015). We tested several hypotheses in the European group, in order to maximize sample size without introducing potential population stratification. Multiple testing correction (FDR) was very conservative; again, we corrected uniformly for the total number of tests. This was to filter out any potentially spurious results. Of the 28 predictions tested, 22 showed effects in the expected direction (p=0.002, one-tailed sign test), and 7 were significant after stringent multiple-

testing correction. Two previous notable null associations, MDD GPS predicting Grade Point Average (GPA), and Type 2 Diabetes GPS predicting GPA, were also null in our sample. Full results are presented in Table 1, including additional associations with the listed GPS phenotypes.

Phenome-Wide Prediction

We also performed hypothesis-free analyses across all 33 GPS and 55 S4S phenotypes to explore potentially novel associations. Multiple prior proportions of causal variants in the genome were tested, as detailed in Methods. Figure 2 presents notable results for GPS prediction of phenotypes in the European group for the prior proportion of 0.3 (that is, an initial assumption that 30% of the genome is associated with the GPS phenotype). The 0.3 prior level showed stronger prediction in past work (Krapohl *et al.* 2015), and corresponds to a plausible assumption about the genetic architecture of many complex traits, due to instances of increasing sample size of GWAS proportionally increasing numbers of associated loci. In this group and prior proportion level, out of 1,815 associations 35 were suggestively significant at $q<0.16$ (the *P*-value threshold corresponding to Akaike Information Criterion (Akaike, 1974)), 11 more were significant at $q<0.05$, and 26 additional associations were robust at $q<0.01$. An additional 53 associations showed at least suggestive significance at other prior levels. A heatmap of analyses at multiple assumed prior proportions of causal variants can be found in Figure 2 (EUR; and for replications in all ancestries, Figures S1-S4 available in the online supplementary materials). Each plot presents significant associations as well as the direction of predictive effect. Because of the stringent correction for multiple testing, we included of interest $q<0.16$ associations, which would be significant with more traditional, liberal correction methods accounting for previously established effects.

Notable unexpected results included SZ GPS significantly predicting nicotine use, depression and anxiety symptoms, and family history of depression, anxiety, alcohol use disorder, and drug use. In addition, GPS for neuroticism (N) predicted a number of relevant psychiatric phenotypes, including neurotic, depression and anxiety symptoms.

Genetic Profile Score Prediction of the Phenome Across Non-European Ancestries

As noted earlier, most discovery GWAS have used European samples, and while there is good evidence for cross-ancestral replication for some traits, the generalizability of many of these relationships transethnically is not known. The diverse ancestry groups within S4S allowed cross-ancestral replication, and the use of continuous GPS metrics made the sample sizes available powerful enough to examine these hypotheses. A large proportion of the

strongest predictors observed in the EUR were replicated across the other ancestries, with a broadly similar pattern of results across all ancestry groups. While some outcome phenotypes were strongly predicted by GPS, a few outcome phenotypes, including physical activity, lifetime history of panic attack, age at first sexual intercourse, and bulimia nervosa were not predicted by any GPS in any ancestry group.

GPS-GPS Correlations

A Priori Hypothesis Testing and Global Cross-Disorder Genetic Profile Analyses

Based on the cross-disorder psychiatric genomics findings to date, we had hypothesized significant GPS correlations between: SZ and BP, SZ and AUT, SZ and MDD, BP and MDD, and AUT and ADHD across each of the ancestry groups. *A priori* hypotheses (described in the Methods and listed in Table 2) of relationships between GPS scores were tested at a prior proportion level of 0.3. Figure 3 presents the results for GPS-GPS partial correlations at a GPS $p = 0.3$, and these results are presented because some phenotypes studied here (e.g., N) were not included in previous analyses. Finally, Figures S5-S8 (available in the online supplementary materials) present these correlations across four non-EUR ancestry groups. Notable unexpected correlations were also observed, including significant positive correlations of neuroticism GPS with GPSs for triglycerides and coronary artery disease. There is some overlap between the discovery samples for neuroticism and triglycerides, but no overlapping studies were included in the neuroticism and the coronary artery disease discovery samples. Therefore, the correlation of neuroticism and coronary artery disease is especially likely to reflect underlying genetic correlation between neuroticism and artery disease. Despite overlap in the discovery samples for the neuroticism and triglycerides polygenic scores, validation using LD score regression supported the existence of a genetic relationship between them ($r_G = 0.53$; $SE = 0.04$; $p = 1.5 \times 10^{-36}$).

Discussion

The findings here present a wide-ranging and nuanced picture of major dimensions of vulnerability to psychopathology at a genetic level. This study includes substantial sample sizes of emerging adults, uses real-world outcome measures (with novel phenotypes in phenomic analyses; see Table S2 in the online supplement for details of assessment scales), includes a wide range of discovery GWAS, and is powerful enough to draw preliminary conclusions about several ancestries. Because this study does not look for “hits” in the traditional GWAS sense and

instead uses continuous GPS metrics, sample sizes provide adequate power across all separate ancestries in this study.

Importantly, results reflect relationships between anxiety, depressive, and schizophrenia-spectrum disorders that are largely consistent with current conceptualizations of diagnostic classification, and confirm the important involvement of a network of medical and risk phenotypes in genetic predisposition to these disorders. Informative genetic associations between medical and clinical phenotypes exist despite the relative dearth of individual loci of genome-wide significance.

We can learn a lot from the study of emerging adults relative to younger, adolescent samples, as more targeted theories about critical exposures and environmental risk factors can emerge. For example, GPS for SZ predicted anxiety, depression, nicotine use, experiences of interpersonal trauma, and family history of mental health problems. Importantly, these results expand on recent evidence that genetic risk for SZ can successfully predict diverse risk phenotypes such as anxiety and negative symptoms (Kendler *et al.* 1996; Fanous *et al.* 2001; Docherty & Sponheim, 2008; Docherty & Sponheim, 2014; Docherty *et al.* 2015; Jones *et al.* 2016; Kendler, 2016), and demonstrate important links between SZ genetic risk and health factors in early adulthood. Significant association of GPS with easily measured, specific risk factors (e.g., nicotine use, family history, trauma) indicates that GPS could be useful in predicting psychopathology, particularly in conjunction with environmental moderators.

The incorporation of personality traits such as N was also quite informative. For example, N GPS significantly predicted a broad network of general anxiety, phobia, panic, N, and depression phenotypes in S4S, as well as multiple health-related GPSs. This is consistent with previous biometrical and genomic research reporting significant relationships of N with MDD (Kendler & Myers, 2010; Genetics of Personality Consortium *et al.* 2015; Docherty *et al.* 2016), and preliminary findings from the UKBiobank suggesting a genetic overlap of neuroticism with cardiovascular health (Gale *et al.* 2016). Conversely, GPS for extraversion predicted decreased depressive symptoms, decreased anxiety symptoms, and decreased family history of mental health problems, though these associations did not remain significant after multiple testing correction. Associations pertaining to GPS for well-being in this sample are forthcoming from our research group.

Notable unexpected GPS-GPS results included positive correlations of N GPS with GPSs for coronary artery disease, which is likely to reflect underlying genetic correlation, as well as with triglycerides. This is the first study we know of to document significant positive genetic associations between N and cardiac health, despite the

high public health cost of N being well-documented (Cuijpers *et al.* 2010; cardiovascular risk and association with psychiatric phenotypes like N may be of special interest to public health efforts). Most of the GPS-GPS *a priori* relationships chosen for replication testing were represented in the same direction across all ancestry groups, corroborating previous efforts to map relationships between genetic risk profiles.

The abundance of significant relationships between intuitive combinations of GPSs and related outcomes is reassuring considering the many factors that could attenuate the statistical link between them. Association between a GPS and an outcome not only reflects correlation between the phenotype in the original ('discovery') GWAS that produced the statistics used to compute the GPS and the outcome phenotype, but is also related to a number of other factors. The link is limited by how accurately the GWAS measured the initial phenotype, how similar the discovery and test samples are (in age, ancestry composition, proportions of each sex, etc.), how well the test phenotype is measured by the data collection instrument, and how well it can incorporate indirect pathways from the genetic architectures to either phenotype.

For example, physical activity increases HDL levels (Kokkinos & Fernhall, 1999), so those who had higher HDL levels in the discovery GWAS (Teslovich *et al.* 2010) were likely a mix of those with innately high levels, those who engaged in higher levels of physical activity, and those with both traits. Therefore, HDL GPS perhaps indexes some propensity to engage in physical activity or other HDL-promoting behaviors, in addition to HDL metabolic variation such as a slower rate of HDL catabolism, which is thought to be the most common genetically determined mechanism of increased HDL levels in humans (Rader, 2006). The portion of the HDL GPS due to fitness behaviors may explain some of the polygenic association with the test phenotypes of BMI and weight.

There are a number of limitations to be aware of when interpreting these results. The synthesis of information from so many sources compounds any methodological and psychometric issues present in the original studies, so there is probable bias in multiple levels of the analysis that is difficult to measure. In order to maintain proximity with real outcomes, we did not transform our phenotypic variables to increase normality, but standardized the continuous variables computed from the participant responses to maintain comparable ranges of measurement. While LDpred performs adequately across ancestry groups, the accuracy in non-European ancestry groups is attenuated to the degree that multiple causal variants fall in regions where LD patterns differ transethnically. In addition, a recent pre-print (Martin *et al.* 2016) shows biased predictions in several different populations using GPS for phenotypes that are also used in this paper (for example, Type 2 diabetes and SZ).

Overall, this broader picture of genetic vulnerability has important implications for how we study risk and resilience in emerging adulthood. While the variance explained by any of these GPSs is small, they provide easily accessible information to guide future prediction, prevention, and intervention efforts to improve health and quality of life outcomes. Future longitudinal and intervention research could elaborate on this atlas to examine the predictive validity and prevention utility of many of the phenotypes here, such as N, family history, trauma, and nicotine use. Future polygenic work would also benefit from GPSs based on non-European ancestry groups when such summary statistics are available. Phenome-wide research utilizing deeper phenotyping methods will likely further enhance results, and thus future prediction of positive and negative health outcomes.

Finally, the relationships outlined here provide implicit suggestions for studies of the causal structure of the GPS phenotypes themselves. The genetic architecture of most of the traits and disorders in the atlas display substantial overlap; a significant portion of genetic variation involved in the etiology of these constructs does not selectively contribute to risk for one phenotype as we know it, but rather has effects that act on some axis of liability that increases the likelihood of many phenotypes. Analyzing multiple related phenotypes in a holistic fashion allows elucidation of the individual patterns of genetic and environmental factors that may explain causal mechanisms—which risk factors they share, and which are unique to one phenotype, thus serving to refine our nosological theories. Any epidemiological analysis is limited if the construct under study is not a uniform disease entity, but as characterization of constructs improves, the power to find their correlates does as well. The better we ask the questions, the more useful the answers become, for both clinical and scientific purposes.

Author Contributions

A. R. Docherty, A. Moscati, and K. S. Kendler developed the study concept. A. Moscati, J. E. Savage, J. E. Salvatore, and M. Cooke contributed to psychometric analyses and data collection. D. Dick and K. S. Kendler oversaw data collection. A. Moscati and A. R. Docherty performed the data analysis and interpretation under the supervision of B. T. Webb, S. A. Bacanu, D. E. Adkins, F. Aliev, A. C. Edwards, and B. P. Riley. A. R. Docherty and A. Moscati drafted the manuscript, and K. Kendler, A. C. Edwards, J. E. Savage, J. E., Salvatore, M. Cooke, B. T. Webb, S. A. Bacanu, D. E. Adkins, A. Moore, R. Peterson, and D. Dick provided edits. All authors approved the final version of the manuscript for submission.

Acknowledgements

We would like to thank the participants and the many VCU faculty, students, and staff who contributed to the design and implementation of this project.

Financial Disclosures

All authors declare no conflict of interest with respect to the authorship or publication of this article. Data collection for the study was funded by R37AA011408, P20AA107828, K02AA018755, and P50AA022537 from the National Institute on Alcohol Abuse and Alcoholism, by Virginia Commonwealth University, and by UL1RR031990 from the National Center for Research Resources and National Institutes of Health Roadmap for Medical Research. A. Docherty was funded by K01MH109765 from the National Institute of Mental Health, and by a Brain & Behavior Research Foundation (formerly NARSAD) Young Investigator Award. A. Moscati, A. Moore were supported by institutional training grant T32MH20030. J. Salvatore was supported by F32AA22269 and K01AA024152. M. Cooke received support from UL1TR000058 from the National Institutes of Health National Center for Advancing Translational Science. A. Edwards was supported by K01AA021399.

Declaration of Interest

Authors report no conflicts of interest.

References

Akaike H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorder (*4th ed.). Washington, DC: American Psychiatric Association

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Benyamin B, Pourcain B, Davis OS, Davies G, Hansell NK, Brion MJ, . . . Visscher PM (2014). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry* **19**, 253-258.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57***,* 289-300.

Brener ND, Kann L, Kinchen SA, Grunbaum JA, Whalen L, Eaton D, . . . Ross JG (2004). Methodology of the youth risk behavior surveillance system. *MMWR Recommendations and Reports* **53**, 1-13.

Brown KW, Ryan RM (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology* **84**, 822-848.

Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI, Schuckit MA (1994). A new, semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *Journal of Studies on Alcohol* **55**, 149-158.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, . . . Neale BM (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47***,* 1236-1241.

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research* **28**, 193-213.

Connor KM, Davidson JR (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety* **18**, 76-82.

Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, . . . Oja P (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine & Science in Sports & Exercise* **35**, 1381-1395.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet* **381***,* 1371-1379.

Cuijpers P, Smit F, Penninx BW, de Graaf R, ten Have M, Beekman AT (2010). Economic costs of neuroticism: A population-based study. *Archives of General Psychiatry* **67,** 1086-1093.

Derogatis LR, Cleary PA (1977). Confirmation of the dimensional structure of the SCL-90: A study in construct validation. *Journal of Clinical Psychology* **33**, 981-989.

Dick DM, Nasim A, Edwards AC, Salvatore JE, Cho SB, Adkins A, . . . Kendler KS (2014). Spit for Science: Launching a longitudinal study of genetic and environmental influences on substance use and emotional health at a large US university. *Fronteirs in Genetics* **5***,* 47.

Docherty AR, Moscati AA, Fanous AH (2016). Cross-disorder psychiatric genomics: A review. *Current Behavioral Neuroscience Reports* **3**, 256-263.

Docherty AR, Moscati A, Peterson R, Edwards AC, Bigdeli TB, Adkins DE, . . . Kendler KS (2016). SNP-based heritability estimates of the personality dimensions and polygenic prediction of both neuroticism and major depression: Findings from CONVERGE. *Translational Psychiatry* **6**, e926.

Docherty AR, Sponheim SR (2008). Anhedonia as a phenotype for the Val158Met COMT polymorphism in relatives of patients with schizophrenia. *Journal of Abnormal Psychology* **117***,* 788-798.

Docherty AR, Sponheim SR (2014). Anhedonia as an indicator of genetic liability to schizophrenia. In M. Ritsner (Ed.), *Anhedonia: A Comprehensive Handbook*. Dordecht: Springer.

Docherty AR, Sponheim SR, Kerns JG (2015). Self-reported affective traits and current affective experiences of biological relatives of people with schizophrenia. *Schizophria Research* **161***,* 340-344.

Fanous A, Gardner C, Walsh D, Kendler KS (2001). Relationship between positive and negative symptoms of schizophrenia and schizotypal symptoms in nonpsychotic relatives. *Archives of General Psychiatry* **58***,* 669-673.

Fisher RA (1919). On the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh: Earth and Environmental Science* **52***,* 399-433.

Gale CR, Hagenaars SP, Davies G, Hill WD, Liewald DC, Cullen B, . . . Harris SE (2016). Pleiotropy between neuroticism and physical and mental health: Findings from 108,038 men and women in UK Biobank. *Translational Psychiatry* **6***,* e791.

Genetics of Personality Consortium, de Moor MH, van den Berg SM, Verweij KJ, Krueger RF, Luciano M, . . . Boomsma DI (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* **72***,* 642-650.

Gray MJ, Litz BT, Hsu JL, Lombardo TW (2004). Psychometric properties of the life events checklist. *Assessment* **11**, 330-341.

Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DC, Ritchie SJ, . . . Deary IJ (2016). Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Molecular Psychiatry* **21***,* 1624-1632.

Hays RD, Sherbourne CD, Mazel RM (1995). *User's manual for the Medical Outcomes Study (MOS): Core measures of health-related quality of life*. Santa Monica, CA: RAND.

Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO (1991). The Fagerstrom Test for Nicotine Dependence: A revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction* **86***,* 1119-1127.

Hendrick SS (1988). A generic measure of relationship satisfaction. *Journal of Marriage and Family* **50**, 93-98.

Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivieres S, Jahanshad N, . . . Medland SE (2015). Common genetic variants influence human subcortical brain structures. *Nature* **520,** 224-229.

Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, . . . Freathy RM (2013): New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature Genetics* **45***,* 76-82.

John OP, Srivastava S (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin (Ed.), *Handbook of Personality: Theory and Research* (pp. 102-183). New York: Guilford Press.

Jones HJ, Stergiakouli E, Tansey KE, Hubbard L, Heron J, Cannon M, . . . Zammit S (2016). Phenotypic manifestation of genetic risk for schizophrenia during adolescence in the general population. *JAMA Psychiatry* **73,** 221-228.

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern D, Hui KY, . . . Cho JH (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124.

Kendler KS (2016). The schizophrenia polygenic risk score: To what does it predispose in adolescence? *JAMA Psychiatry* **73***,* 193-194.

Kendler KS, Gardner CO, Prescott CA (1997). Religion, psychopathology, and substance use and abuse; A multimeasure, genetic-epidemiologic study. *American Journal of Psychiatry* **154**, 322-329.

Kendler KS, Jacobson K, Myers JM, Eaves LJ (2008). A genetically informative developmental study of the relationship between conduct disorder and peer deviance in males. *Psychologial Medicine* **38**, 1001-1011.

Kendler KS, Karkowski LM, Prescott CA (1999) Causal relationship between stressful life events and the onset of major depression. *American Journal of Psychiatry* **156**, 837-841.

Kendler KS, Karkowski-Shuman L, Walsh D (1996). The risk for psychiatric illness in siblings of schizophrenics: The impact of psychotic and non-psychotic affective illness and alcoholism in parents. *Acta Psychiatrica Scandinavica* **94***,* 49-55.

Kendler KS, Myers J (2010). The genetic and environmental relationship between major depression and the five-factor model of personality. *Psychological Medicine* **40**, 801-806.

Kendler KS, Prescott CA (1999). Caffeine intake, tolerance, and withdrawal in women: A population-based twin study. *American Journal of Psychiatry* **156**, 223-228.

Kessler RC, Angermeyer M, Anthony JC, de Graaf R, Demyttenaere K, Gasquet I, . . . Ustun TB (2007): Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* **6**, 168-176.

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* **62***,* 593-602.

Kokkinos PF, Fernhall B (1999). Physical activity and high density lipoprotein cholesterol levels: What is the relationship? *Sports Medicine* **28**, 307-314.

Krapohl E, Euesden J, Zabaneh D, Pingault JB, Rimfeld K, von Stumm S, . . . Plomin R (2015). Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry* **21***,* 1188-1193.

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, . . . Amouyel P (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452-1458.

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, . . . Speliotes EK (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206.

Lynam DR, Smith GT, Whiteside SP, Cyders MA (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior. Technical report.* West Lafayette, IN: Purdue University.

Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* **18**, 497-511.

Marigorta UM, Navarro A (2013): High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genetics* **9***,* e1003566.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Grave S, Daly MJ, . . . Kenny EE (2016). Population genetic history and polygenic risk biases in 1000 genomes. Unpublished manuscript. Available at: http://biorxiv.org/content/early/2016/08/23/070797

Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, . . . McCarthy MI (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981-990.

Neale BM, Medland SE, Ripke S, Asherson P, Franke B, Lesch KP, . . . Nelson, S (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* **49***,* 884-897.

Okbay A, Baselmans BM, de Neve JE, Turley P, Nivard MG, Fontana MA, . . . Cesarini D (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624-633.

Otowa T, Hek K, Lee M, Byrne EM, Mirza SS, Nivard MG, . . . Hettema JM (2016) Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular Psychiatry* **21**, 1391-1399.

Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, . . . Ong KK (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* **43**, 977-983.

Rader DJ (2006). Molecular regulation of HDL metabolism and function: Implications for novel therapies. *The Journal of Clinical Investigation* **116**, 3090-3100.

Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, . . . Koellinger PD (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340***,* 1467-1471.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427.

Schuckit MA, Smith TL, Tipp JE (1997). The Self-Rating of the Effects of alcohol (SRE) form as a retrospective measure of the risk for alcoholism. *Addiction* **92**, 979-988.

Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, . . . Samani NJ (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333-338.

Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, . . . Mohlke KL (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, . . . Plenge RM (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42**, 508-514.

Steinberg L, Lamborn SD, Dornbusch SM, Darling N (1992). Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child Development* **63**, 1266-1281.

Taal HR, St Pourcain B, Thiering E, Das S, Mook-Kanamori DO, Warrington NM, . . . Early Growth Genetics Consortium (2012). Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nature Genetics* **44**, 532-538.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, . . . Kathiresan S (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713.

Tobacco and Genetics Consortium (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics* **42**, 441-447.

van der Valk RJ, Kreiner-Moller E, Kooijman MN, Guxens M, Stergiakouli E, Saaf A, . . . Jaddoe VW (2015). A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics* **24**, 1155-1168.

van den Berg SM, de Moor MH, Verweij KJ, Krueger RF, Luciano M, Arias Vasquez A, . . . Boomsma DI (2016). Meta-analysis of genome-wide association studies for extraversion: Findings from the genetics of personality consortium. *Behavior Genetics* **46***,* 170-182.

Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, . . . Price AL (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics* **97**, 576-592.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Frayling TM (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173-1186.

Table 1. Phenotype Prediction at .3 GPS: Tests of *A Priori* Associations Based on Previous Research

Discovery Phenotype	Test Phenotype	Previous Finding	Sign of Coefficient	Other Observed Associations
Age at Menarche	BMI/Obesity	- (Bulik-Sullivan et al., 2015)	-***	-Weight*, +Extraversion*
	Height	+ (Bulik-Sullivan et al., 2015)	+*	
Alzheimer's Disease	GPA	- Educational Attainment (Hagenaars et al., 2016)	+	None
Anorexia	BMI/Obesity	- (Bulik-Sullivan et al., 2015)	-	None
Attention Deficit Disorder	GPA	- Educational Attainment (Hagenaars et al., 2016)	-	-Parental Education*
Autism	GPA	+ Educational Attainment (Hagenaars et al., 2016)	-	+Parental Education*
Bipolar Disorder	Depression	+ MDD (Bulik-Sullivan et al., 2015)	+	-BMI**, -Weight***, +Family History for Alcohol*, Drug use**, and Depression/Anxiety**
	GPA	+ Educational Attainment (Hagenaars et al., 2016)	+	
Birth Height	Height	+ (Bulik-Sullivan et al., 2015)	+***	None
Birth Weight	Height	+ (Bulik-Sullivan et al., 2015)	+	None
Body Mass Index	BMI	+ (Krapohl et al., 2015)	+***	+Weight***, +Subjective Response to Ethanol***, -GAD Screener*
	GPA	- Educational Attainment (Hagenaars et al., 2016)	-	
Child IQ	GPA	+ Cognitive Ability (Krapohl et al., 2015)	+	None
College	BMI/Obesity	- (Bulik-Sullivan et al., 2015)	-	-Subjective Response to Ethanol*, -Conscientiousness*, +Parental Education***
	Cigarette Use	- Ever smoker (Bulik-Sullivan et al., 2015)	+	
	GPA	+ Cognitive Ability (Krapohl et al., 2015)	+*	
Coronary Artery Disease	GPA	- Educational Attainment (Hagenaars et al., 2016)	-	None
	Height	-	-	
Ever Smoker	BMI/Obesity	+ (Bulik-Sullivan et al., 2015)	+	+Antisocial High School Behavior*
Height	GPA	+ Educational Attainment (Hagenaars et al., 2016)	-	-BMI**, +Weight***
	Height	+ (Krapohl et al., 2015)	+***	
Infant Head Circumference	Height	+ (Bulik-Sullivan et al., 2015)	+	None
Intracranial Volume	GPA	+ Educational Attainment (Hagenaars et al., 2016)	-	None
Major Depressive Disorder	GPA	Null (Hagenaars et al., 2016)	Null	+Depressive symptoms**, +Neuroticism*, +GAD Screener*, +Specific Phobia Screener*, +Family History for Alcohol***, and Depression/Anxiety***
Schizophrenia	Depression	+ MDD (Bulik-Sullivan et al., 2015)	+**	+Alcohol Use Disorder Symptoms*. +Anxiety Symptoms***, +Neuroticism*, +Interpersonal Trauma***, +PTSD Screener*, +Ever Use Nicotine*, +GAD Screener*, +Social Phobia Screener*, +Family History of Alcohol*, Drug Use**, and Depression/Anxiety***
	Cigarette Use	Null, Ever Smoker (Bulik-Sullivan et al., 2015)	+**	
	GPA	+ Educational Attainment (Hagenaars et al., 2016)	-	
Type 2 Diabetes	GPA	Null (Hagenaars et al., 2016)	Null	None

Note: * = q<0.16, ** = q<0.05, *** = q<0.01. BMI = body mass index; GPA = grade point average; GAD = generalized anxiety disorder; MDD = major depressive disorder; PTSD = posttraumatic stress disorder; S4S = Spit for Science.

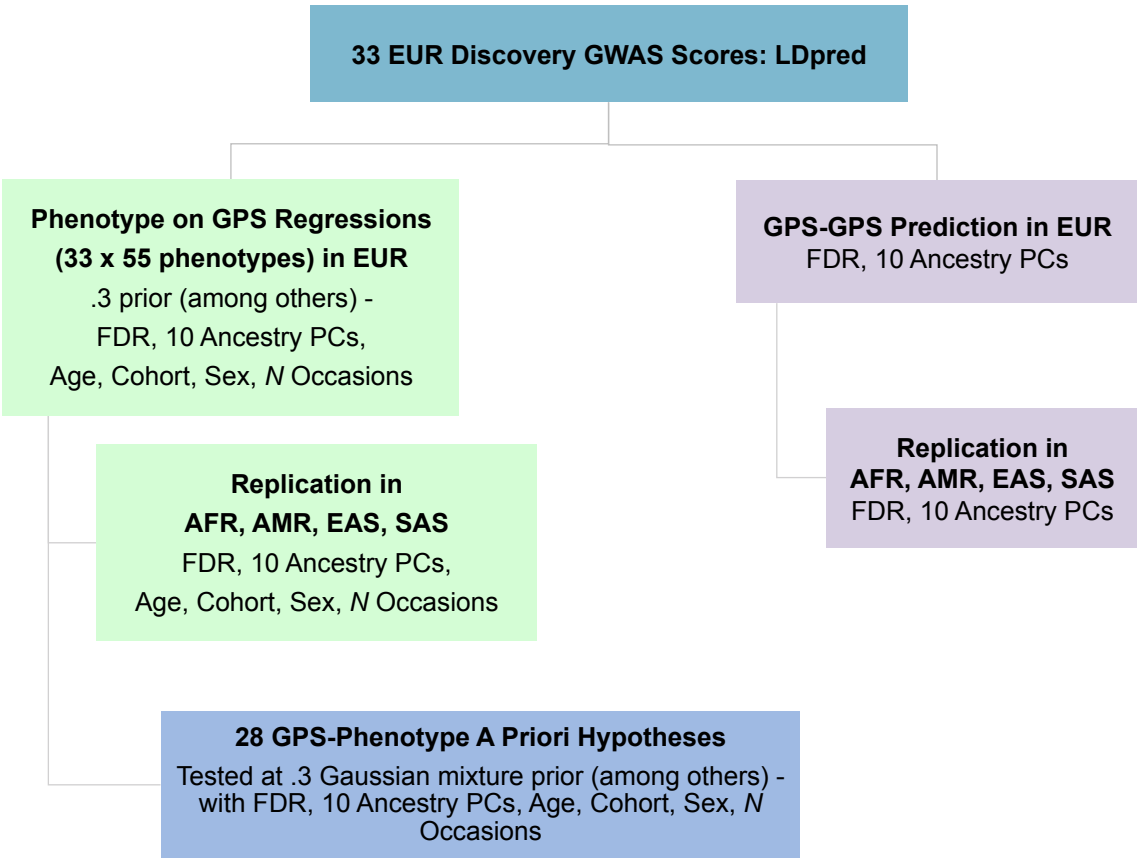
Table 2. Cross-Ancestry Replication Tests of European GPS-GPS Correlations Based on Atlas from Bulik-Sullivan et al., 2015

Discovery Phenotype	Previous Finding (Bulik-Sullivan et al., 2015)	Significant Replication in Spit for Science
Age at Menarche	+Height, +HDL	+Height**** +HDL**
Alzheimer's	-College	-College*
Anorexia	+Schizophrenia	+Schizophrenia****
Bipolar Disorder	+College, +Schizophrenia, +Major Depression	+College****, +Schizophrenia****, +Major Depression****
Birth Weight	-Type 2 Diabetes, +Infant Head Circumference	-Type 2 Diabetes* (+ in EAS and SAS), +Infant Head Circumference****
College	-Ever Smoked, -Triglycerides, -Coronary Artery Disease, +HDL	-Ever Smoked*, -Triglycerides****, -Coronary Artery Disease**** (+ in AMR and EAS), +HDL****
Coronary Artery Disease	+Type 2 Diabetes, +Triglycerides	+Type 2 Diabetes****, +Triglycerides****
Crohn's	+Ulcerative Colitis	+Ulcerative Colitis****
HDL	-Type 2 Diabetes, -Waist Hip Ratio, -Triglycerides	-Type 2 Diabetes****, -Waist Hip Ratio****, -Triglycerides****
Height	+Infant Head Circumference	+Infant Head Circumference****
LDL	+Triglycerides	+Triglycerides****
Major Depressive Disorder	+Schizophrenia	+Schizophrenia****

Note: **** = q-value < 0.0001, *** = q-value < 0.001, ** = q-value < 0.01, * = q-value < 0.05.

522
523
524

Figures



525

526 **Fig. 1. Flow Chart of the GPS-Phenome and GPS-GPS Analyses.** GWAS = genome wide association study;

527 FDR = false discovery rate; GPS = genome-wide polygenic score; PC = principal component.

528

529

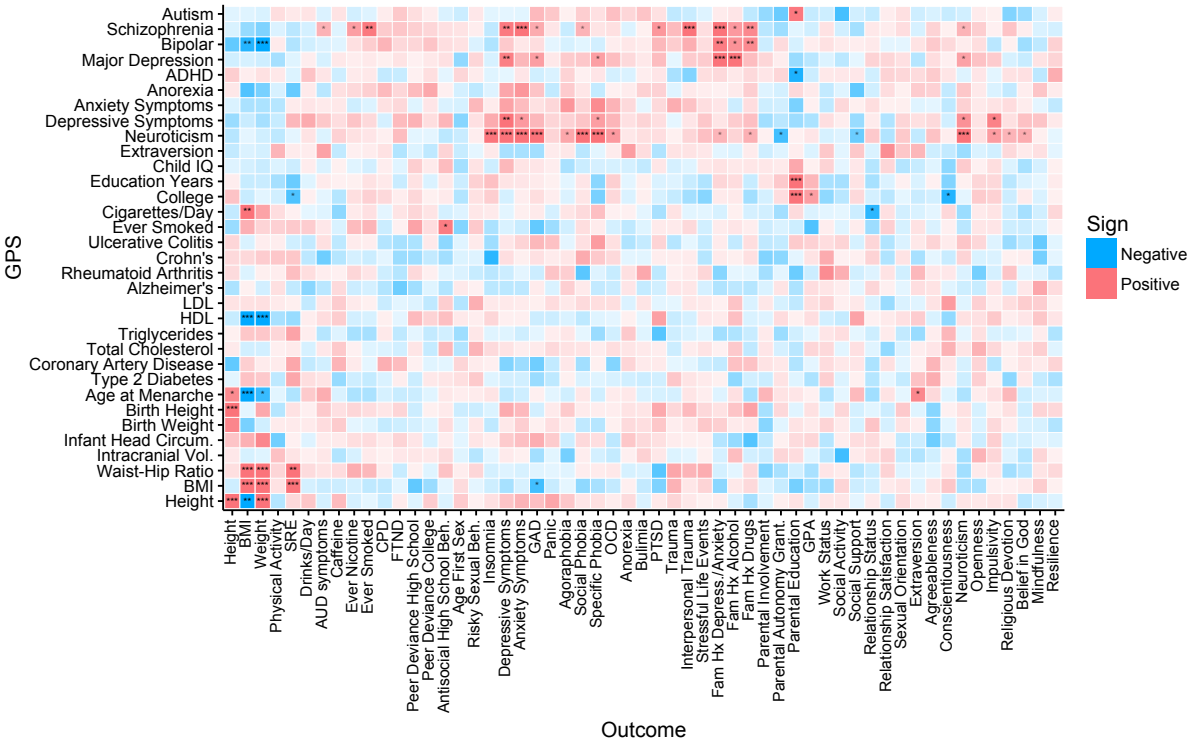


Fig. 2. GPS on phenome regression q-values at GPS with prior proportion of causal effects = 0.3. Here, asterisks in the cells of the heatmap denote results of greater effect: *** = q-value < .01, ** = q-value < .05, * = q-value < .16. Blue values reflect a negative association, and red reflect positive association. Intensity of color indicates $-\log_{10}$ p-value.

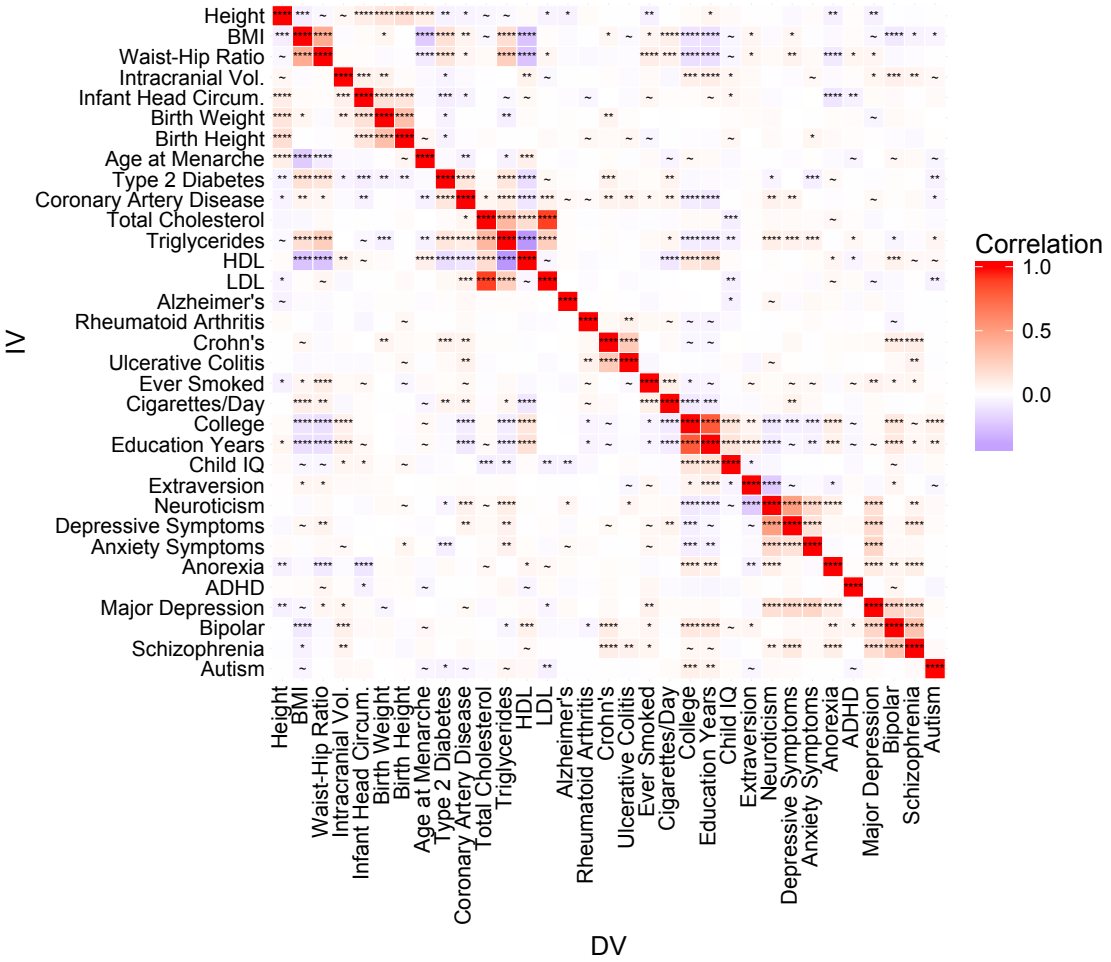


Fig. 3. Genetic Overlap and Co-Heritability of GPS in European Sample (EUR). Heatmap of partial correlation coefficients between GPS with prior proportion of causal effects = 0.3. Here, asterisks in the cells of the heatmap denote results of greater effect: **** = q-value < 0.0001, *** = q-value < 0.001, ** = q-value < 0.01, * = q-value < 0.05, and ~ = suggestive significance at q-value <0.16. Blue values reflect a negative correlation, and red reflect positive correlation.