Genetics of educational attainment aid in identifying biological subcategories of schizophrenia

Authors: Vikas Bansal[1,2,3], Marina Mitjans[1,2,4], Casper A.P. Burik[5,6], Richard Karlsson Linnér[5,6], Aysu Okbay[5,6], Cornelius A. Rietveld[6,7], Martin Begemann[2,4,8], Stefan Bonn[3,4], Stephan Ripke[9,10,11], Michel G. Nivard[12,13], Hannelore Ehrenreich[2,4,13], Philipp D. Koellinger[5,6,13]

Affiliations:

1 These co-authors contributed equally.

2 Clinical Neuroscience, Max Planck Institute of Experimental Medicine, Hermann-Rein-Straße 3, 37075, Göttingen, Germany

3 Research Group for Computational Systems Biology, German Center for Neurodegenerative Diseases (DZNE), Von-Siebold-Straße 3A, 37075 Göttingen, Germany

4 DFG Research Center for Nanoscale Microscopy and Molecular Physiology of the Brain (CNMPB), Humboldtallee 23, 30703, Göttingen, Germany

5 Department of Complex Trait Genetics, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV, Amsterdam, Netherlands

6 Erasmus University Rotterdam Institute for Behavior and Biology, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, Netherlands

7 Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, Netherlands

8 Department of Psychiatry & Psychotherapy, University of Göttingen, Von-Siebold-Straße 5, 37075, Göttingen, Germany

9 Analytic and Translational Genetics Unit, Massachusetts General Hospital, 02114 MA, Boston, USA

10 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, 02142 MA, Cambridge, USA

11 Department of Psychiatry and Psychotherapy, Charité-Universitätsmedizin Berlin, Campus Mitte, Berlin, 10117, Germany

12 Department of Biological Psychology, Vrije Universiteit Amsterdam, van der Boechorststraat 1, 1081 BT, Amsterdam, Netherlands

13 These authors jointly directed this work.

Correspondence: Professor Philipp D. Koellinger, Complex Trait Genetics, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV, Amsterdam, Netherlands, E-mail: p.d.koellinger@vu.nl

ABSTRACT

Higher educational attainment (EA) is known to have a protective effect regarding the severity of schizophrenia (SZ). However, recent studies have found a small positive genetic correlation between EA and SZ. Here, we investigate possible causes of this counterintuitive finding using genome-wide association results for EA and SZ (n = 443,581) and a replication cohort (1,169 controls and 1,067 cases) with high-quality SZ phenotypes. We find strong genetic overlap between EA and SZ that cannot be explained by chance, linkage disequilibrium, or assortative mating. Instead, our results suggest that the current clinical diagnosis of SZ comprises at least two disease subtypes with non-identical symptoms and genetic architectures: One part resembles bipolar disorder (BIP) and high intelligence, while the other part is a cognitive disorder that is independent of BIP.

INTRODUCTION

Schizophrenia (SZ) is the collective term used for a severe, highly heterogeneous and costly psychiatric disorder that is caused by a complex interplay of environmental and genetic factors[1–4]. The latest genome-wide association study (GWAS) by the Psychiatric Genomics Consortium (PGC) identified 108 genomic loci that are associated with SZ[5]. These 108 loci jointly account for ≈3.4% of the variation on the liability scale to SZ[5], while all single nucleotide polymorphisms (SNPs) that are currently measured by SNP arrays capture ≈64% (s.e. = 8%) in the variation in liability to the disease[6]. This suggests that most of the genetic variants contributing to the heritability of SZ have very small effects and that they have not been isolated yet. This could be due in part to the fact that the clinical disease classification of SZ spans across many different syndromes (e.g., catatonia, paranoia, grandiosity, difficulty in abstract thinking, thought blocking, social withdrawal, hallucinations) that may not have identical genetic architectures. Therefore, identifying additional genetic variants and understanding through which pathways they influence specific SZ syndromes is an important step in understanding the etiologies of the 'schizophrenias'[7]. However, GWAS analyses of specific SZ syndromes would require very large sample sizes to be statistically well-powered, and the currently available datasets on deeply phenotyped SZ individuals are not large enough yet for this purpose.

Here, we use an alternative approach that combines data for SZ with another cognitive phenotype that can be studied in very large GWAS samples—educational attainment (EA). The relationship between SZ and EA is peculiar: There are contradictory results on the relationship between SZ and EA from phenotypic and genetic data that can be used as an avenue to further our understanding about SZ. Phenotypic data seem to suggest a *negative* correlation between EA and SZ[8]. For example, SZ patients with lower EA typically show an earlier age of disease onset, higher levels of psychotic symptomatology, and worsened global cognitive function[8]. In fact, EA has been suggested to be a measure of premorbid function and predictor of outcomes in SZ. Moreover, it has been forcefully argued that retarded intellectual development during childhood and bad school performance should be seen as core features of SZ and early indicators of the disease that precede the development of psychotic symptoms[9,10]. Furthermore, credible genetic links between SZ and impaired cognitive performance have been found[11].

In contrast to these findings, recent studies using GWAS results identified a small, but *positive genetic* correlation between EA and SZ[12,13]. Here, we explore possible reasons for this contradictory result using the largest, non-overlapping GWAS samples on cognitive traits to date, totaling 443,581 individuals of European descent (the vast majority of observations coming from EA). For follow-up analyses, we use data from an independent replication

sample that has exceptionally detailed measures of SZ symptoms, the GRAS (Göttingen Research Association for Schizophrenia) data collection[4,7,14].

As a first step, we used the proxy-phenotype method (PPM) to illustrate the genetic overlap between EA and SZ. As a side-result, this approach may isolate novel empirically plausible candidate genes for SZ, comparable to similar studies using PPM that have demonstrated this for cognitive performance[15], Alzheimer's disease, intracranial and hippocampal volume[13], depression and neuroticism[16]. PPM is a two-stage approach that increases statistical power by using genetic association results from a large, independent sample for a related phenotype to limit the multiple testing burden for the phenotype of interest[15]. Previous evidence suggests a strong genetic overlap between EA and SZ, which implies that EA could be used as a proxy-phenotype for SZ because EA can be studied in much larger samples[13,16]. However, compared to the present work, these previous studies used substantially smaller and partially overlapping samples and did not have access to an independent cohort that could be used for replication and follow-up analyses.

There are several possible reasons why EA-associated SNPs may also be associated with SZ. One possibility is that a set of genes that is generally important for all brain-related phenotypes is driving this enrichment. This hypothesis suggests that the set of genetic loci that our proxy-phenotype analysis identifies should be generally enriched for association with all brain-related phenotypes, but not for non-brain-related outcomes. To investigate this possibility, we test genetic loci that are jointly associated with EA and SZ for enrichment across 21 additional traits (**Supplementary Note**).

Second, enrichment could also be a generic consequence of EA-associated SNPs exhibiting above average linkage disequilibrium (LD) with neighboring SNPs. This would increase the probability that these SNPs "tag" other genetic variants that are associated with SZ, or any other disorder[12]. To investigate this possibility, we propose a measure that tests for enrichment beyond what is expected for each EA related SNP given its LD with its neighbors (**Supplementary Note**).

A third possible cause of strong enrichment and weak genetic correlation is heterogeneity in SZ—i.e., sub-types of the disease having different biological causes and varying genetic correlations with EA. Heterogeneity in the disease may also be a reason why previous studies did not succeed in predicting specific syndromes of SZ using a "normal" polygenic score (PGS) that was derived from large-scale GWAS on SZ, which implicitly assumed that all SZ-associated SNPs influence all syndromes in the same way[4,17]. If heterogeneity in the disease is causing the observed enrichment of EA with SZ, the sign concordance pattern of SNPs with both traits may contain relevant information that is pertinent to specific SZ syndromes. We tested this by constructing PGS in our replication cohort with high-quality SZ phenotypes that take the sign concordance of SNPs for EA and SZ into account (**Supplementary Note**). As a robustness check, we repeat this analysis excluding patients diagnosed with schizoaffective disorder.

A fourth possible cause of enrichment is that other phenotypes are genetically correlated with both EA and SZ. Previous studies indicated a particularly strong positive genetic correlation between SZ and bipolar disorder (BIP), which may influence the genetic overlap of both diseases with related phenotypes such as EA, childhood intelligence (IQ), and neuroticism[12,13,18]. We use genome-wide inferred statistics (GWIS) that allow controlling for the genetic correlation between SZ and BIP to investigate how "unique" SZ (controlling for BIP) and "unique" BIP (controlling for SZ) are related to EA, childhood IQ, and neuroticism[18].

A fifth possible cause may be assortative mating, which has been demonstrated both for EA[19] and SZ[20]. We use simulations to explore if independent assortative mating for the two phenotypes may induce a spurious genetic overlap.

This list of potential causes for the genetic overlap between EA and SZ may not be exhaustive and several of these factors may be at work simultaneously.

RESULTS

Proxy-phenotype analyses

Figure 1 presents an overview of the proxy-phenotype analyses. The first-stage GWAS on EA (**Supplementary Note**) identified 506 loci that passed our predefined threshold of $P_{EA} < 10^{-5}$ (https://osf.io/dnhfk/); 108 of them were genome-wide significant ($P_{EA} < 5 \times 10^{-8}$, see **Supplementary Table 5.1**). Of the 506 EA lead-SNPs, 132 are associated with SZ at nominal significance ($P_{SZ} < 0.05$), and 21 of these survive Bonferroni correction ($P_{SZ} < \frac{0.05}{506} = 9.88 \times 10^{-5}$). LD score regression results suggest that the vast majority of the association signal in both the EA[13] and the SZ[5] GWAS are truly genetic signals, rather than spurious signals originating from uncontrolled population stratification. **Figure 2a** shows a Manhattan plot for the GWAS on EA highlighting SNPs that were also significantly associated with SZ (red crosses for $P_{SZ} < 0.05$, green crosses for $P_{SZ} = 9.88 \times 10^{-5}$).

A Q-Q plot of the 506 EA lead SNPs for SZ is shown in **Figure 2b**. Although the observed sign concordance of 52% is not significantly different from a random pattern ($P = 0.40$), we find 3.23 times more SNPs in this set of 506 SNPs that are nominally significant for SZ than expected given the distribution of the *P* values in the SZ GWAS results (raw enrichment $P = 6.87 \times 10^{-10}$, **Supplementary Note**). The observed enrichment of the 21 EA lead SNPs that pass Bonferroni correction for SZ ($P_{SZ} < \frac{0.05}{506} = 9.88 \times 10^{-5}$) is even more pronounced (27 times stronger, $P = 5.44 \times 10^{-14}$).

Bayesian credibility of the results

The effect sizes of these 21 SNPs on SZ are small, ranging from *Odds* = 1.02 (rs4500960) to *Odds* = 1.11 (rs4378243) after winner's curse correction (**Table 1**). However, Bayesian calculations with reasonable prior beliefs (e.g., 1% or 5%, **Supplementary Note**) suggest that most of these 21 SNPs are likely or virtually certain to be truly associated with SZ.

Prediction of future genome-wide significant loci for schizophrenia

Of the 21 variants we identified, 12 are in LD with loci previously reported by the PGC[5] and 2 are in the major histocompatibility complex (MHC) region on chromosome 6 and were therefore not separately reported in that study. Three of the variants we isolated (rs7610856, rs143283559, rs28360516) were independently found in a recent meta-analysis of the PGC results[5] with another large-scale sample[21]. We show in the **Supplementary Note** that using EA as a proxy-phenotype for SZ helped to predict the novel genome-wide significant findings reported in that study, which illustrates the power of the proxy-phenotype approach. Furthermore, two of the 21 variants (rs756912, rs7593947) are in LD with loci recently reported in a study that also compared GWAS findings from EA and SZ using smaller samples and a less conservative statistical approach[22] (**Supplementary Note**). The remaining 2 SNPs we identified (rs7336518 on chr13 and rs7522116 on chr 1) add to the list of empirically plausible candidate loci for SZ.

LD-aware enrichment across different traits

Figure 3 and **Supplementary Table 5.2** show the LD-aware enrichment of the SNPs that are jointly associated with EA and SZ across 22 traits. We find significant joint LD-aware enrichment of this set of SNPs for SZ, BIP, neuroticism and childhood IQ, and for inflammatory bowel disease and age at menarche. However, we find no LD-aware enrichment for other brain-traits that are phenotypically related to SZ, such as depressive symptoms, subjective well-being, autism, and attention deficit hyperactivity disorder. We also do not find LD-aware enrichment for most traits that are less obviously related to the brain (e.g., BMI, coronary artery disease) and our negative controls (e.g., fasting insulin, birth weight, birth length). Furthermore, one of the novel SNPs we isolated shows significant LD-aware enrichment both for SZ and for BIP (rs7522116).

Replication in the GRAS sample

A PGS based on the 132 loci jointly associated with both EA and SZ (*SZ_132*) adds $\Delta R^2 = 7.54\% - 7.01\% = 0.53\%$ predictive accuracy for the SZ case-control status to a PGS (*SZ_all*) derived from the GWAS on SZ alone ($P = 1.7 \times 10^{-4}$, **Table 2**, Model 3). The *SZ_132* score also significantly adds ($P = 3.4 \times 10^{-4}$) to the predictive accuracy of the SZ case-control status when all other scores we constructed are included as control variables. In addition to *SZ_132*, PGS for SZ (*SZ_all*) and for BIP (*BIP_all*) also predict case-control status, jointly reaching an adjusted ΔR^2 of ≈ 9% (**Table 2**, Model 9 and **Supplementary Note**).

Polygenic prediction of schizophrenia measures in the GRAS sample

We find that the number of years of education is phenotypically correlated with later age at prodrome, later onset of disease, and less severe disease symptoms among SZ patients in the GRAS sample (**Supplementary Note**, **Supplementary Table 8.1 and Supplementary Fig. 1**). The *EA_all* score is associated with years of education ($P = 2.6 \times 10^{-6}$) and premorbid IQ ($P = 2.3 \times 10^{-4}$) among SZ patients (**Supplementary Note** and **Table 3**). Consistent with earlier results[4], we find that none of the SZ measures can be predicted by the normal SZ PGS (*SZ_all*, **Supplementary Table 8.2**). Importantly, by utilizing GWAS results from both EA and SZ, we show that it is possible to predict specific features of SZ (Global Assessment of Functioning (GAF), Clinical Global Impression of Severity (CGI-S), and Positive and Negative Syndrome Scale (PANSS)) from genetic data. In a multiple regression analysis[23] that allows a "ceteris paribus" interpretation of the included variables, we find that the *EA_all* score is associated with less severe disease outcomes only if we condition on the effects of the *Concordant* and *Discordant* scores. And conditional on the *EA_all* score, the *Concordant* and *Discordant* scores are associated with more (less) severe positive and negative symptoms as measured by the PANSS scale, respectively (**Table 3**). The best predictive accuracy of SZ readouts using these scores is currently observed for GAF (R^2 = 1.38%). Of note, several of the symptoms measured by PANSS are also symptoms of BIP. The degree and composition of symptoms varies with the phase at evaluation (manic or depressive) and the general disease severity. We repeated these analyses excluding patients who were diagnosed with schizoaffective disorder (SD) and found similar results, implying that the genetic heterogeneity in SZ that we identify is not only due to SD (**Supplementary Note**, **Supplementary Table 8.4.a**).

Controlling for the genetic overlap between schizophrenia and bipolar disorder

None of the EA-associated lead SNPs ($P_{EA} < 5 \times 10^{-8}$) are significantly associated with "unique" $SZ_{(\text{min BIP})}$ after Bonferroni correction (**Supplementary Table 9.1**, **Supplementary Note**). The sign concordance of the EA lead SNPs with "unique" $SZ_{(\text{min BIP})}$ was 44.5% ($P =$

0.046). **Supplementary Figure 2** shows a Q-Q plot of the EA lead-SNPs for "unique" $SZ_{(min\ BIP)}$. Although we find 1.6 times more EA-associated SNPs with $P_{SZunique} < 0.05$ than expected by chance (raw enrichment $P = 0.02$, **Supplementary Note**), the enrichment is much weaker than in the main SZ GWAS results that did not control for the genetic overlap between SZ and BIP. The genetic correlations between EA $SZ_{(min\ BIP)}$, and IQ and $SZ_{(min\ BIP)}$ are negative and significant (r_g = -0.16, $P = 3.88\times10^{-04}$ and r_g = -0.31, $P = 6.00\times10^{-03}$ respectively), which is in line with the idea of SZ being a cognitive disorder[9]. Furthermore, the genetic correlations of EA and IQ with $BIP_{(min\ SZ)}$ remain positive and get somewhat stronger (r_g = 0.31, $P = 2.87\times10^{-07}$ and r_g = 0.33, $P = 3.18\times10^{-02}$ respectively) compared with the ordinary BIP GWAS results. However, controlling for the genetic overlap of SZ and BIP does not affect the genetic correlations with neuroticism (**Figure 4**).

Simulations of assortative mating

Our simulations for assortative mating were based on relatively extreme assumptions that increased our chance of finding spurious enrichment of EA loci for SZ. The results suggest it is unlikely that assortative mating is a major cause for the genetic overlap we observe between EA and SZ (**Supplementary Fig. 3**).

Biological annotations

Biological annotation of the 132 SNPs that are jointly associated with EA and SZ using DEPICT identified 111 significant reconstituted gene sets (**Supplementary Table 10.1**). Pruning these resulted in 19 representative gene sets including dendrites, axon guidance, transmission across chemical synapses, and abnormal cerebral cortex morphology (**Supplementary Table 10.2** and **Figure 5a**). All significantly enriched tissues are related to the nervous system and sense organs (**Figure 5b**). Furthermore, "Neural Stem Cells" is the only significantly enriched cell-type (**Supplementary Table 10.3**). DEPICT prioritized genes that are known to be involved in neurogenesis and synapse formation (**Supplementary Table 10.4**). Some of the genes, including *SEMA6D* and *CSPG5*, have been suggested to play a potential role in SZ[24,25]. For the two novel candidate SNPs reported in this study (rs7522116 and rs7336518), DEPICT points to the *FOXO6* (Forkhead Box O6) and the *SLITRK1* (SLIT and NTRK Like Family Member 1) genes, respectively. *FOXO6* is predominantly expressed in the hippocampus and has been suggested to be involved in memory consolidation, emotion and synaptic function[26,27]. Similarly, *SLITRK1* is also highly expressed in the brain[28], particularly localized to excitatory synapses and promoting their development[29], and it has previously been suggested to be a candidate gene for neuropsychiatric disorders[30].

DISCUSSION

We explored the genetic overlap between EA and SZ using the largest currently available GWAS sample on human cognitive traits to date. Using EA as a proxy-phenotype, we identified 21 genetic loci for SZ and showed that this approach helps to predict future GWAS hits for SZ. We isolated two additional candidate genes for SZ, *FOXO6* and *SLITRK1*. Our results show that EA-associated SNPs are much more likely to also be associated with SZ than expected by chance. However, these genetic loci do not influence both traits with a systematic sign pattern that would correspond to a strong positive or negative genetic correlation.

The results of our follow-up analyses are most consistent with two hypotheses that complement each other: First, the genetic overlap between EA and SZ is to some extent induced by pleiotropic effects of many genes that affect not only EA and SZ but also other

traits such as BIP and IQ. Second, different syndromes of SZ (e.g., low cognitive performance and psychosis) seem to be driven by different genetic effects. The clinical diagnosis of SZ aggregates over these different syndromes. In particular, our results suggest that the current clinical diagnosis of SZ comprises at least two disease subtypes with non-identical symptomatology and genetic architectures: One part resembles bipolar disorder (BIP) and high intelligence, while the other part is a cognitive disorder that is independent of BIP. Consistent with this idea, we find that PGS that take the sign concordance of SNPs with EA and SZ into account begin for the first time to predict specific SZ features from genetic data (R^2 between 0.4% and 1.4%), while this was not possible with "ordinary" PGS for SZ.

Other mechanisms that we explored, in particular LD-patterns of the EA-associated SNPs and assortative mating, do not seem to be major drivers of the genetic overlap between EA and SZ. Furthermore, the loci we identified in our PPM analysis do not seem to be associated with all brain-related phenotypes, suggesting some degree of phenotype-specificity of the results. We note that the enrichment for age at menarche of the SNPs that are jointly associated with EA and SZ may be related to the final stage of brain development which coincides with the onset of puberty[31–34].

The highly complex genetic architecture of the "schizophrenias" that our results point to implies that most patients will have individual-specific genetic loads for either subtype of the disease, contributing to individual differences in symptoms. The genetic heterogeneity we identified could imply that treatments will vary in their effectiveness across disease subtypes.

Overall, our study corroborates that EA is a useful proxy-phenotype for psychiatric outcomes. Specifically, combining GWAS results from EA and SZ led to the identification of two seemingly distinct subcategories of SZ. Even though each of them may still harbor highly heterogeneous disease subgroups, the new subcategories can pave the way for further biological subgroup analyses. Therefore, a psychiatric nosology that is based on biological causes rather than pure phenotypical classifications may be feasible in the future. Studies that combine well-powered GWAS from several diseases and from phenotypes that represent variation in the normal range such as EA are likely to play an important part in this development. However, deep phenotyping of large patient samples will be inevitable to link GWAS results from complex outcomes such as EA and SZ to specific biological disease subgroups.

AUTHOR CONTRIBUTIONS

P.D.K. designed and oversaw the study and conducted proxy-phenotype analyses. V.B. and M.M. carried out analyses in the GRAS sample. V.B. conducted bioinformatics and computed the LD-aware enrichment tests, which were developed by M.N. C.A.P.B. conducted simulation analyses. M.N. computed GWIS results and genetic correlations. R.K.L. assisted with biological annotation and visualization of results. P.D.K., V.B., M.M., and H.E made especially major contributions to writing and editing. All authors contributed to and critically reviewed the manuscript.

ACKNOWLEDGMENTS

This research was carried out under the auspices of the Social Science Genetic Association Consortium (SSGAC), including use of the UK Biobank Resource. We thank all research consortia that provide access to GWAS summary statistics in the public domain. Specifically, we acknowledge data access from the Psychiatric Genomics Consortium (PGC), the Genetic Investigation of ANthropometric Traits Consortium (GIANT), the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), the International Genomics of Alzheimer's Project (IGAP), the CARDIoGRAMplusC4D Consortium, the Reproductive Genetics Consortium (ReproGen), the Tobacco and Genetics Consortium (TAG), the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC), the ENIGMA Consortium, and the Childhood Intelligence Consortium (CHIC). We would like to thank the customers and employees of 23andMe for making this work possible as well as Joyce J. Tung, Nick. A. Furlotte, and David. A Hinds from the 23andMe research team. This study was supported by funding from an ERC Consolidator Grant (647648 EdGe, Philipp D Koellinger), the Max Planck Society, the Max Planck Förderstiftung, the DFG (CNMPB), EXTRABRAIN EU-FP7, the Niedersachsen-Research Network on Neuroinfectiology (N-RENNT), and EU-AIMS. Michel G Nivard was supported by Royal Netherlands Academy of Science Professor Award to Dorret I Boomsma (PAH/6635). Additional acknowledgements are provided in the Supplementary Online Materials.

COMPETING FINANCIAL INTERESTS

The authors declare no conflict of interests.

REFERENCES

1 Knapp M, Mangalore R, Simon J. The global costs of schizophrenia. *Schizophr Bull* 2004; **30**: 279–293.

2 Sullivan PF, Kendler KS, Neale MC, KS K, SB T, DB P *et al.* Schizophrenia as a complex trait. *Arch Gen Psychiatry* 2003; **60**: 1187.

3 Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* 2015; **47**: 702–709.

4 Stepniak B, Papiol S, Hammer C, Ramin A, Everts S, Hennig L *et al.* Accumulated environmental risk determining age at schizophrenia onset: a deep phenotyping-based study. *The Lancet Psychiatry* 2014; **1**: 444–453.

5 Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**: 421–427.

6 Bhatia G, Gusev A, Loh P, Vilhjálmsson BJ, Ripke S, PGC *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases. 2015http://dx.doi.org/10.1101/022418.

7 Ehrenreich, H; Mitjans, M; Van der Auwera, S; Centeno, TP; Begemann, M; Grabe, HJ; Bonn, S; Nave K-A. OTTO: a new strategy to extract mental disease-relevant combinations of GWAS hits from individuals. *Mol Psychiatry* 2017. doi:10.1038/mp.2016.208.

8 Swanson CL, Gur RC, Bilker W, Petty RG, Gur RE. Premorbid educational attainment in schizophrenia: association with symptoms, functioning, and neurobehavioral measures. *Biol Psychiatry* 1998; **44**: 739–747.

9 Kahn RS, Keefe RSE, JD H, B E, GM K, H D *et al.* Schizophrenia is a cognitive illness. *JAMA Psychiatry* 2013; **70**: 1107.

10 Kraepelin E. *Psychiatrie: Ein Lehrbuch für Studierende und Ärzte*. 4th ed. Verlag von Johann Ambrosius Barth: Leipzig, Germany, 1893.

11 Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 2013; **505**: 361–366.

12 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**: 1236–1241.

13 Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**: 539–542.

14 Ribbe K, Friedrichs H, Begemann M, Grube S, Papiol S, Kästner A *et al.* The cross-sectional GRAS sample: A comprehensive phenotypical data collection of schizophrenic patients. *BMC Psychiatry* 2010; **10**: 91.

15 Rietveld CA, Esko TT, Davies G, Pers TH, Turley PA, Benyamin B *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci U S A* 2014; **111**: 13790–13794.

16 Okbay A, Baselmans BML, Neve J-E De, Turley P, Nivard MG, Fontana MA *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* doi:10.1038/ng.3552.

17 Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet Part B Neuropsychiatr Genet* 2015; **168**: 649–659.

18 Nieuwboer HA, Pool R, Dolan CV, Boomsma DI, Nivard MG. GWIS: Genome-wide inferred statistics for functions of multiple phenotypes. *Am J Hum Genet* 2016; **99**: 917–927.

19 Hugh-Jones D, Verweij KJH, St. Pourcain B, Abdellaoui A. Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence* 2016; **59**: 103–108.

20 Nordsletten AE, Larsson H, Crowley JJ, Almqvist C, Lichtenstein P, Mataix-Cols D *et al.* Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* 2016; **73**: 354.

21 Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. 2016http://dx.doi.org/10.1101/068593.

22 Le Hellard S, Wang Y, Witoelar A, Zuber V, Bettella F, Hugdahl K *et al.* Identification of gene loci that overlap between schizophrenia and educational attainment. *Schizophr Bull* 2016. doi:10.1093/schbul/sbw085.

23 Wooldridge JM. Multiple Regression Analysis: Estimation. In: *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2013, pp 70–76.

24 So H-C, Fong PY, Chen RYL, Hui TCK, Ng MYM, Cherny SS *et al.* Identification of neuroglycan C and interacting partners as potential susceptibility genes for schizophrenia in a Southern Chinese population. *Am J Med Genet Part B Neuropsychiatr Genet* 2010; **153B**: 103–113.

25 Arion D, Horváth S, Lewis DA, Mirnics K. Infragranular gene expression disturbances in the prefrontal cortex in schizophrenia: signature of altered neural development? *Neurobiol Dis* 2010; **37**: 738–46.

26 Salih DAM, Rashid AJ, Colas D, de la Torre-Ubieta L, Zhu RP, Morgan AA *et al.* FoxO6 regulates memory consolidation and synaptic function. *Genes Dev* 2012; **26**: 2780–2801.

27 Maiese K. FoxO Proteins in the Nervous System. *Anal Cell Pathol* 2015; **2015**: 1–15.

28 Aruga J, Yokota N, Mikoshiba K. Human SLITRK family genes: genomic organization and expression profiling in normal brain and brain tumor tissue. *Gene* 2003; **315**: 87–94.

29 Beaubien F, Raja R, Kennedy TE, Fournier AE, Cloutier J-F, Ichtchenko K *et al.* Slitrk1 is localized to excitatory synapses and promotes their development. *Sci Rep* 2016; **6**: 27343.

30 Proenca CC, Gao KP, Shmelkov S V, Rafii S, Lee FS, Aruga J *et al.* Slitrks as emerging candidate genes involved in neuropsychiatric disorders. *Trends Neurosci* 2011; **34**: 143–53.

31 Huttenlocher, R. P. Synapse elimination and plasticity in developing human cerebral

cortex. *Am J Ment Defic* 1984; **88**: 488–496.

32 Purves D, Lichtman J. Elimination of synapses in the developing nervous system. *Science (80-)* 1980; **210**: 153–157.

33 Yazici E, Bursalioglu FS, Aydin N, Yazici AB. Menarche, puberty and psychiatric disorders. *Gynecol Endocrinol* 2013; **29**: 1055–1058.

34 Saugstad LF. Age at puberty and mental illness. Towards a neurodevelopmental aetiology of Kraepelin's endogenous psychoses. *Br J Psychiatry* 1989; **155**: 536–544.

35 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.

36 Sheskin D. The binomial sign test for a single sample. In: *Handbook of Parametric and Nonparametric Statistical Procedures*. Taylor & Francis Group: Boca Raton, 2007, pp 289–311.

37 Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–295.

38 Consortium C-DG of the PG. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013; **381**: 1371–1379.

39 Pers TH, Karjalainen JM, Chan Y, Westra H-JH-J, Wood AR, Yang J *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 2015; **6**: 5890.

40 Frey BJ, Dueck D. Clustering by passing messages between data points. *Science (80-)* 2007; **315**.

ONLINE METHODS

All reported statistical results are based on two-sided tests, unless indicated otherwise. Our proxy-phenotype analyses and our replication strategy followed a pre-registered analysis plan (https://osf.io/dnhfk/). The full description of all materials and methods is provided in the **Supplementary Note**.

GWAS on educational attainment.

The EA sample excluded all cohorts that participated in the GWAS on SZ described below, yielding a sample size of n = 363,502 individuals of European descent[13]. The GRAS replication sample was not part of the GWAS on EA, either.

GWAS on schizophrenia.

The SZ sample consisted of n = 34,409 cases and n = 45,670 controls, diagnosed with SZ or schizoaffective disorder[5]. We excluded the GRAS data collection from the GWAS on SZ.

Proxy-phenotype look-up.

Analyses were carried out using 8,240,280 autosomal SNPs that passed quality controls in both GWAS and additional filters described in the **Supplementary Note**. We selected approximately independent lead SNPs from the EA GWAS results using the clumping procedure in PLINK[35]. We looked up the SZ results of all approximately independent EA lead-SNPs that passed the pre-defined significance threshold of $P_{EA} < 10^{-5}$.

We tested if the observed sign concordance between EA and SZ is different from 50% using the binomial probability test[36]. "Raw" enrichment factors and "raw" enrichment p-values of the EA lead-SNPs on SZ were calculated by taking the actual distribution of P values in the SZ GWAS result files into account but ignoring the LD scores[12,37].

LD-aware enrichment across different traits.

We developed an enrichment test that corrects for the LD score of each SNP (**Supplementary Note**). We conducted this test for the 132 SNPs that are jointly associated with EA and SZ in our proxy-phenotype analyses ($P_{EA} < 10^{-5}$ and $P_{SZ} < 0.05$). LD scores were obtained from the HapMap 3 European reference panel. We investigated SZ and 21 additional traits for which GWAS results were available in the public domain. Some of the traits were chosen because they are phenotypically related to SZ (e.g., BIP), while others were less obviously related to SZ (e.g., age at menarche) or served as negative controls (e.g., fasting insulin). If one of the 132 candidate SNP was not available in the reference panel or the GWAS results of the other traits, we tried to use a good proxy, yielding 79 to 105 available SNPs per trait.

Phenotypic correlations.

We explored the correlations between the number of years of education with 7 quantitative measures of SZ in the GRAS sample of SZ cases: Age at prodrome, age at disease onset, premorbid IQ, GAF, CGI-S, and PANSS positive and PANSS negative scores.

Replication and Bayesian credibility of results.

Our replication uses a PGS in the GRAS data collection, which is based on the 132 independent EA lead-SNPs that are also nominally associated with SZ ($P_{EA} < 10^{-5}$ and $P_{SZ} < 0.05$). This PGS (called *SZ_132*) was constructed using the regression coefficient estimates of the SZ GWAS as weights. In addition to this polygenic replication strategy, we further probed the credibility of our results using a heuristic Bayesian calculation.

Polygenic prediction of schizophrenia symptoms in the GRAS sample.

We predicted the number of years of education and 7 quantitative measures of SZ in the GRAS sample of SZ cases. For each phenotype, we separately compared the predictive performance of several PGS: Scores constructed from the full GWAS result on SZ, EA, BIP, and neuroticism (called *SZ_all*, *EA_all*, *BIP_all*, *Neuro_all*, respectively); scores constructed using only the 132 SNPs that are jointly associated with EA and SZ (called *EA_132* and *SZ_132*, using EA and SZ GWAS coefficients as weights, respectively); and two scores that split the *SZ_all* score into two parts based on sets of SNPs that either have concordant or discordant effects on EA and SZ (called *Concordant* and *Discordant*). Genetic outliers of self-reported non-European descent (n = 13 cases) were excluded from the analysis.

Controlling for the genetic overlap between schizophrenia and bipolar disorder.

We estimated GWIS[18] to obtain SNP regression coefficients that are unique to SZ, which are corrected for the genetic overlap between SZ and BIP. The SZ samples used in the GWIS are not overlapping with the samples used in the EA GWAS and they exclude our replication sample (GRAS). BIP GWAS results were obtained from the PGC[38]. We refer to the set of obtained summary statistics as "unique" $SZ_{(\text{min BIP})}$. We then repeated the look-up of the EA-associated lead SNPs in those summary statistics as described above. Similarly, we obtained GWIS results for "unique" $BIP_{(\text{min SZ})}$ using the same method and data. We computed genetic correlations of these GWIS results with EA, childhood intelligence, and neuroticism using bivariate LD score regression[12] and compared the results to those obtained using ordinary SZ and BIP GWAS results.

Simulations of assortative mating.

We conducted simulations to test if strong assortative mating on EA and SZ can induce a spurious genetic overlap between the two traits.

Biological annotation.

To gain first insights into possible biological pathways that are indicated by the genetic loci identified by our PPM analysis, we applied DEPICT[13,39] using a false discovery rate (FDR) threshold of ≤ 0.05. To identify independent biological groupings, we used the Affinity Propagation method based on the Pearson distance matrix for clustering[40].

Data availability.

GWAS meta-analysis results for EA and SZ as well as GWIS results for unique" $SZ_{(\text{min BIP})}$ and "unique" $BIP_{(\text{min SZ})}$ can be downloaded from the SSGAC website (https://www.thessgac.org/). For information about the GRAS data collection, contact the principal investigator of the study: Hannelore Ehrenreich (ehrenreich@em.mpg.de).

Code availability.

Computer code used to generate LD-aware enrichment and GWIS results can be downloaded from the SSGAC website (https://www.thessgac.org/).

Table 1: SNPs significantly associated with schizophrenia after Bonferroni correction.

	SNP-ID	R^2 (adj)	*Odds* (adj)	*EAF*	Power ($\alpha = 0.05/506$)	Posterior probability of true association, Prior belief (π) 0.1%	1.0%	5.0%	10.0%
1	rs79210963	0.021%	0.931	0.89	22.9%	75.0%	96.8%	99.3%	99.7%
2	rs7610856	0.022%	0.955	0.41	22.8%	74.9%	96.8%	99.3%	99.7%
3	rs10896636	0.020%	0.956	0.67	17.8%	68.7%	95.6%	99.1%	99.5%
4	rs756912	0.022%	0.956	0.51	22.7%	74.8%	96.7%	99.3%	99.7%
5	rs6449503	0.020%	0.961	0.51	12.9%	60.0%	93.7%	98.7%	99.3%
6	**rs7336518**	**0.014%**	**0.964**	**0.13**	**1.5%**	**13.4%**	**60.6%**	**88.5%**	**93.9%**
7	rs143283559	0.017%	0.965	0.72	4.6%	32.8%	83.0%	96.1%	98.0%
8	rs11210935	0.014%	0.973	0.77	1.2%	10.9%	55.1%	86.0%	92.5%
9	rs77000541	0.018%	0.974	0.33	1.6%	14.1%	62.2%	89.2%	94.3%
10	rs2819344	0.017%	0.983	0.62	0.3%	3.0%	23.3%	60.4%	75.3%
11	rs4500960	0.017%	1.017	0.47	0.3%	3.0%	23.3%	60.4%	75.3%
12	rs28360516	0.013%	1.027	0.70	1.4%	12.6%	59.0%	87.8%	93.5%
13	**rs7522116**	**0.015%**	**1.029**	**0.56**	**3.0%**	**23.8%**	**75.8%**	**94.0%**	**96.9%**
14	rs7593947	0.018%	1.040	0.51	12.5%	59.1%	93.5%	98.6%	99.3%
15	rs11694989	0.021%	1.044	0.43	17.9%	68.8%	95.7%	99.1%	99.5%
16	rs320700	0.024%	1.054	0.65	36.4%	85.3%	98.3%	99.7%	99.8%
17	rs3957165	0.020%	1.056	0.83	14.7%	63.6%	94.6%	98.9%	99.4%
18	rs10791106	0.026%	1.056	0.54	46.9%	89.9%	98.9%	99.8%	99.9%
19	rs2992632	0.025%	1.060	0.74	36.8%	85.5%	98.3%	99.7%	99.8%
20	rs10773002	0.043%	1.087	0.28	91.0%	99.0%	99.9%	100.0%	100.0%
21	rs4378243	0.044%	1.112	0.85	91.5%	99.1%	99.9%	100.0%	100.0%

Notes: The SNPs in the table are order by their *Odds* ratio on schizophrenia. Effect sizes for schizophrenia (in R^2 and *Odds*) are downward adjusted for the winner's curse. R^2 was approximated from the winner's curse adjusted *Odds* ratios, using the formulas described in **Supplementary Note**. The winner's curse adjustment took into account that only SNPs with $P = 0.05/506$ were selected. Power calculations assumed that the available GWAS sample size for schizophrenia for each SNP consisted of 34,409 cases and 45,670 controls. EAF is the effect allele frequency in the schizophrenia GWAS data. SNPs highlighted in bold are associations for schizophrenia that have not been emphasized in the previous literature.

Table 2: Polygenic prediction of schizophrenia status in the GRAS sample.

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
SZ_132	*standardized beta*	0.11**		0.08**						0.07**
	P value	5.4×10^{-08}		1.7×10^{-04}						3.4×10^{-04}
SZ_all	*standardized beta*		0.31**	0.30**						0.27**
	P value		6.1×10^{-38}	1.5×10^{-34}						2.6×10^{-28}
EA_132	*standardized beta*				-3.7×10^{-03}		-0.01			-0.03
	P value				0.86		0.66			0.14
EA_all	*standardized beta*					0.04	0.04			0.04
	P value					0.08	0.07			0.08
BIP_all	*standardized beta*							0.18**		0.13**
	P value							$1.\ 4\times10^{-17}$		1.4×10^{-09}
Neuro_all	*standardized beta*								0.03	0.02
	P value								0.17	0.27
	n	2,223	2,223	2,223	2,223	2,223	2,223	2,223	2,223	2,223
	ΔR^2	0.0125	0.0701	0.0754	-0.0004	0.0009	0.0006	0.0312	0.0004	0.0914

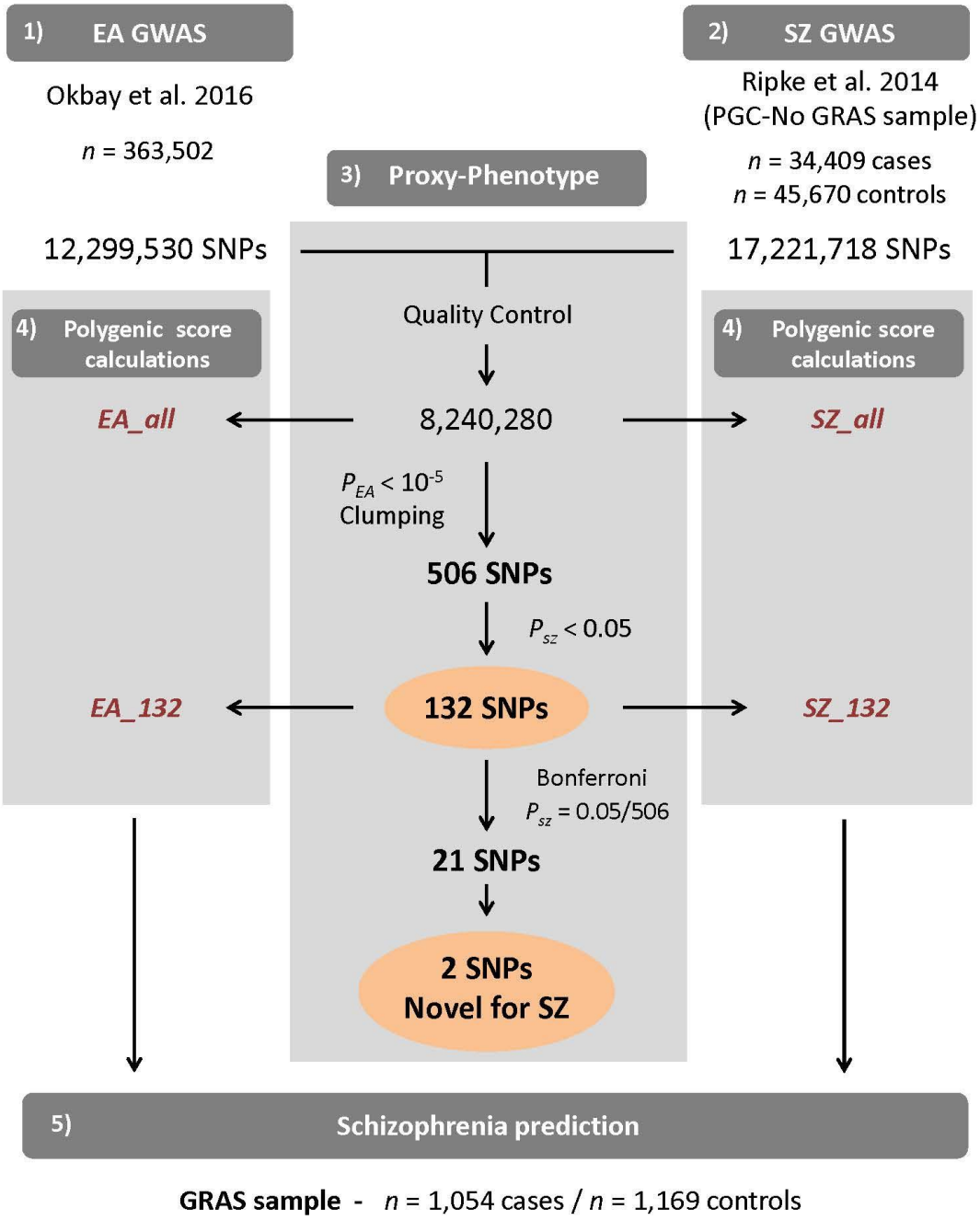
Notes: The reported effects are the standardized beta values of linear probability model (LPM) for schizophrenia status. Models 1 – 9 differed only in the inclusion of the variables displayed in this table. All models also included the first 10 genetic principal components as control variables. *denotes significance at $P < 0.05$. ** denotes significance at $P < 0.001$. The ΔR^2 is the difference between the adjusted R^2 of the model compared to the baseline model that only included control variables but no polygenic scores.

Table 3: Polygenic risk prediction of schizophrenia outcomes in the GRAS sample.

		Years of education[1]	Age at prodrome	Age at disease onset	Premorbid IQ[1]	GAF[2]	CGI-S[2]	PANSS positive[2]	PANSS negative[2]
Concordant	*standardized beta*	-0.05	-0.01	-0.04	-0.07	-0.16*	0.10*	0.13*	0.20**
	P value	0.31	0.86	0.45	0.19	2×10^{-3}	0.05	0.01	2.5×10^{-04}
Discordant	*standardized beta*	0.04	-0.03	-0.02	0.01	0.13*	-0.07	-0.10*	-0.16*
	P value	0.47	0.57	0.71	0.90	0.01	0.18	0.05	2.8×10^{-03}
EA_all	*standardized beta*	0.21**	7.5×10^{-04}	0.01	0.17**	0.16**	-0.11*	-0.07	-0.17**
	P value	2.6×10^{-06}	0.99	0.85	2.3×10^{-04}	2.6×10^{-04}	0.01	0.09	2.4×10^{-04}
BIP_all	*standardized beta*	0.07*	-0.03	-0.03	0.06	-6.6×10^{-04}	0.03	0.02	-0.03
	P value	0.02	0.35	0.34	0.10	0.98	0.37	0.52	0.36
Neuro_all	*standardized beta*	-0.04	0.05	0.05	0.01	-0.07*	0.03	0.04	-0.01
	P value	0.20	0.11	0.13	0.81	0.03	0.31	0.18	0.85
	n	1,039	915	1,043	903	1,010	1,014	1,009	1,002
	ΔR^2	0.0350	-0.0003	0.0008	0.0225	0.0138	0.0043	0.0037	0.0120

Notes: Linear regression using the first 10 genetic principal components as control variables. [1]: Age of onset was included as covariate. [2]: Medication was included as covariate. *denotes significance at $P < 0.05$. ** denotes significance after Bonferroni correction ($P < 0.05/40 = 0.00125$). The ΔR^2 is the difference between the adjusted R^2 of the model compared to the baseline model that only included control variables but no polygenic scores.

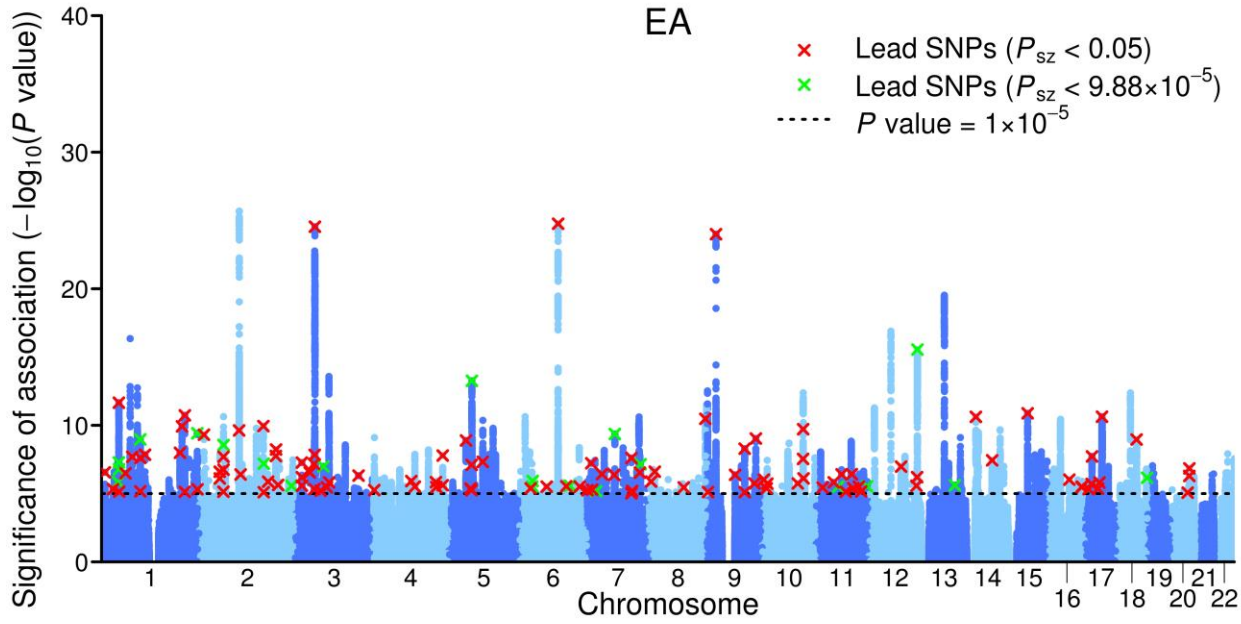
Figure 1: Workflow of the proxy-phenotype analyses



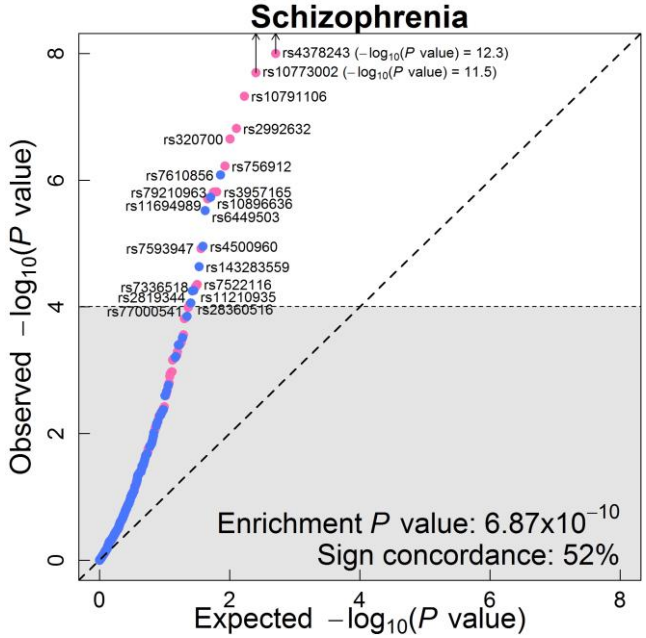
Notes: Educational attainment (EA) and schizophrenia (SZ) GWAS results are based on the analyses reported in ref. [5,13]. All cohorts that were part of the SZ GWAS were excluded from the meta-analysis on EA. The GRAS data collection was not included in either the SZ or the EA meta-analysis. Proxy-phenotype analyses were conducted using 8,240,280 autosomal SNPs that passed quality control. Genetic outliers of non-European descent (n = 13 cases) were excluded from the analysis in the GRAS data collection (**Supplementary Note**).

Figure 2: Results of the proxy-phenotype analyses.

a. Manhattan plot for educational attainment (EA) associations (*n* = 363,502).



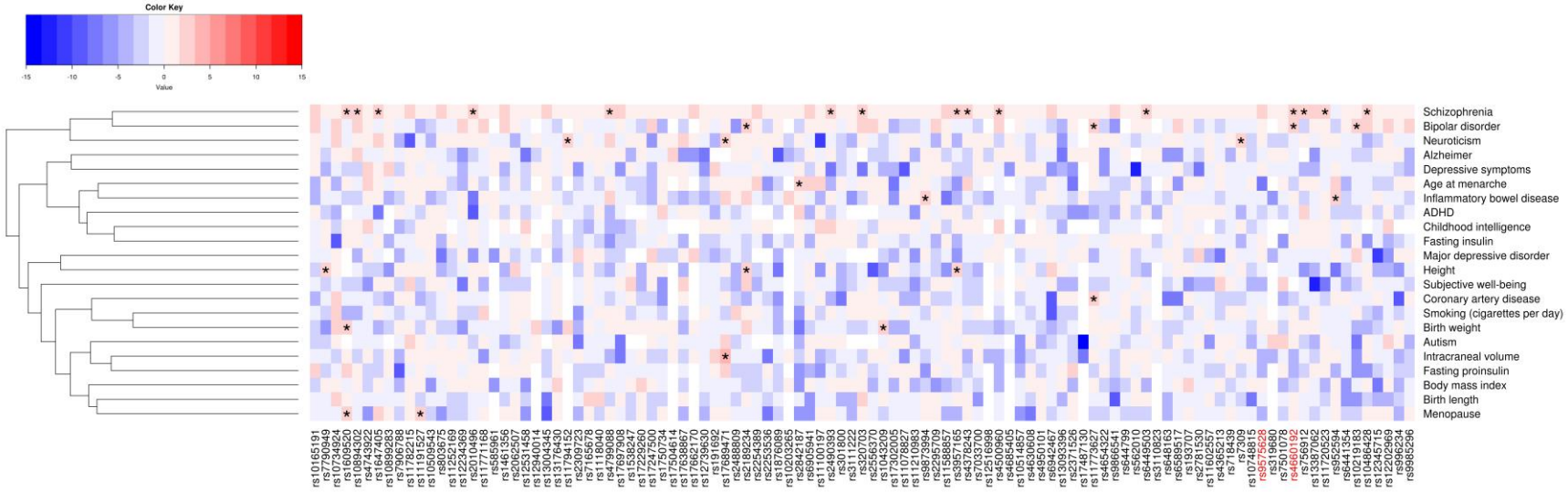
b. Q–Q plot of the 506 EA-associated SNPs for schizophrenia (SZ) (*n* = 34,409 cases and *n* = 45,670 controls).



Notes: Panel a: The *x* axis is the chromosomal position, and the *y* axis is the significance on the $-\log_{10}$ scale. The black dashed line shows the suggestive significance level of 10^{-5} that we specified in our preregistered analysis plan. Red and green crosses identify EA-associated lead-SNPs that are also associated with SZ at nominal or Bonferroni-adjusted significance levels, respectively.

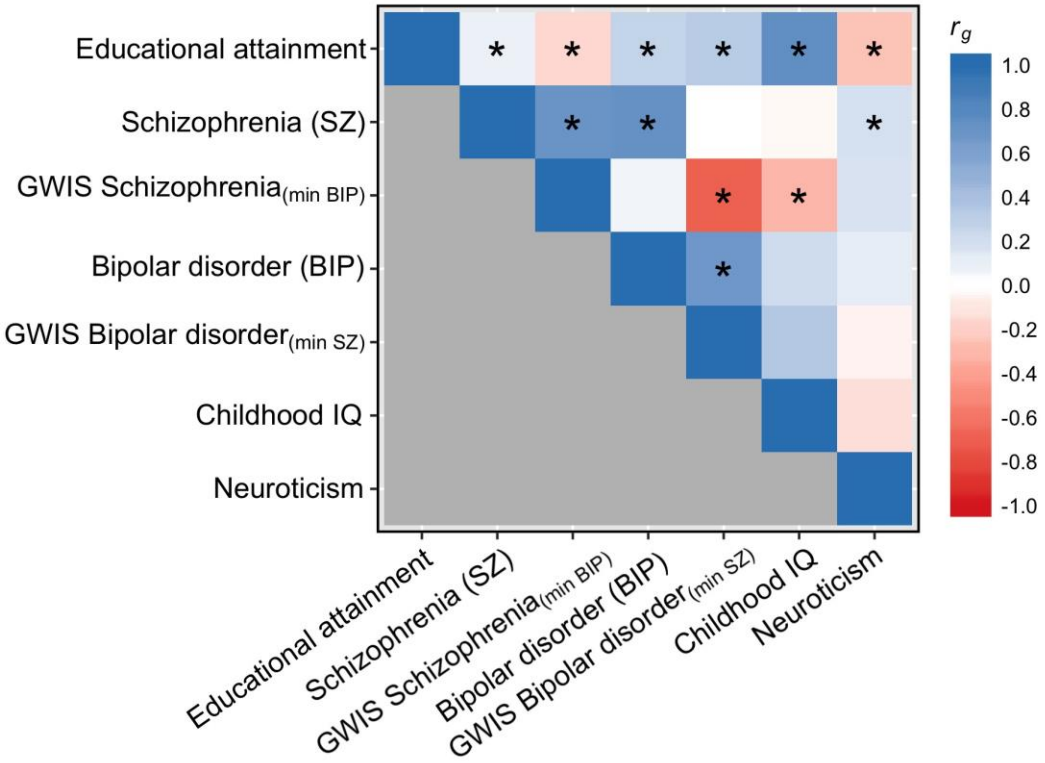
Panel b: SNPs with concordant effects on both phenotypes are pink, and SNPs with discordant effects are blue. SNPs outside the grey area (21 SNPs) pass the Bonferroni-corrected significance threshold that corrects for the total number of SNPs we tested ($P < 0.05/506 = 9.88\times10^{-5}$) and are labelled with their rs numbers. Observed and expected *P* values are on the $-\log_{10}$ scale. For the sign concordance test: $P = 0.40$.

Figure 3: LD-aware enrichment across traits for SNPs that are jointly associated with EA ($P_{EA} < 10^{-5}$) and SZ ($P_{SZ} < 0.05$).



Notes: Color codes illustrate the degree of LD-aware enrichment. Red and blue represent stronger and weaker LD-aware enrichment than expected, respectively. Darker colors illustrate more substantial deviation from expectation. A star (*) indicates that the observed LD-aware enrichment is significant after Bonferroni correction for the number of SNPs that were tested for the specific trait (ranging from 79 to 105). Note that the *p*-values of this test are not the same as those reported in the GWAS and proxy-phenotype analyses because the hypothesis that was test here is different (**Supplementary Note**). The lines on the left side represent hierarchical clustering using the euclidean distance and the complete agglomeration method. Note that we restricted our analysis to HapMap3 SNPs because these are available in most publically available GWAS summary statistics. If a candidate SNP from our proxy-phenotype results was not included in HapMap3, we replaced it by the best available proxy SNP that was available ($LD_{r^2} > 0.8$ and a maximum distance of 500 kb to our missing EA lead SNPs, choosing the proxy with the highest LD_{r^2}). See **Supplementary Note** for more details.

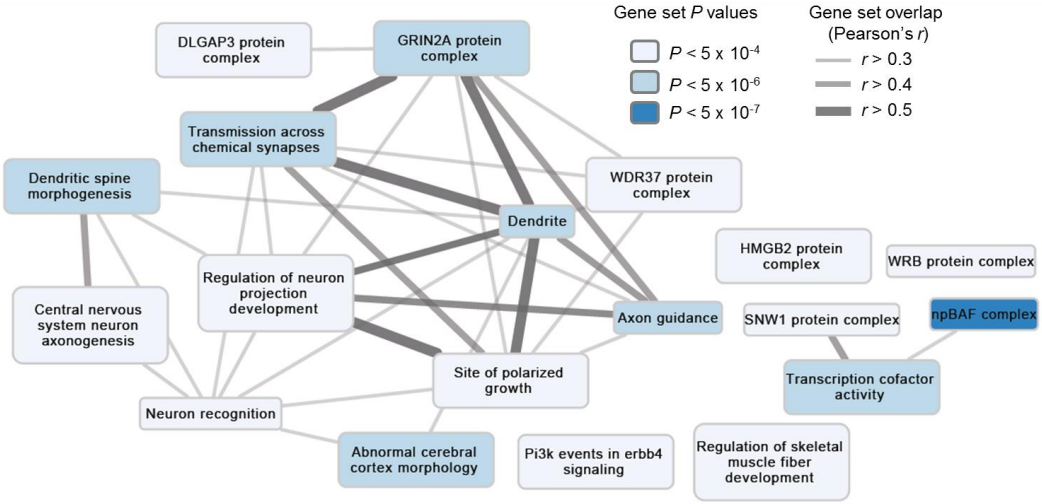
Figure 4: Genetic correlations of GWAS and GWIS results that are central to the relationship between SZ and EA.



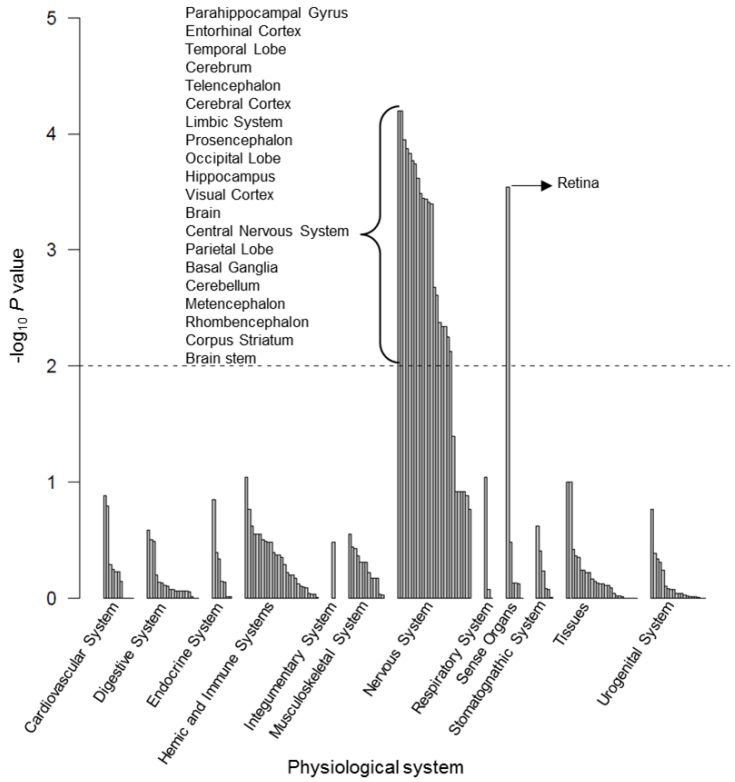
Notes: The heatmap displays the genetic correlations across 7 sets of GWAS or GWIS summary statistics. Genetic correlations were estimated with LD score regression[12]. The color scale represents the genetic correlations ranging from –1 (red) to 1 (blue). Asterisk denotes that the genetic correlation is significant at *P* value < 0.01, without adjustment for multiple comparisons.

Figure 5: Biological annotation of SNPs that are jointly associated with EA ($P_{EA} < 10^{-5}$) and SZ ($P_{SZ} < 0.05$).

a. Gene set enrichment using DEPICT



b. Tissue enrichment using DEPICT



Notes: Panel a: Exemplar gene sets identified using DEPICT at FDR < 5%. Color represents the DEPICT gene set enrichment *P* value without adjustment for multiple comparisons (lower *P* values are reflected by darker colors). Gene sets with Pearson correlations above $r = 0.3$ are connected by edges. Panel b: Tissue enrichment results obtained by DEPICT at FDR < 5%. Names of the 21 significantly enriched tissues are shown above the dashed line. *P* values are without adjustment for multiple comparison. The dotted line represents the 0.01 *P* value. See **Supplementary Note** for additional details.

Supplementary Note to accompany "Genetics of educational attainment aid in identifying biological subcategories of schizophrenia"

Table of contents

1 GWAS on educational attainment

We obtained GWAS summary statistics on educational attainment (EA) from the Social Science Genetic Association Consortium (SSGAC). The results are based on the analyses reported in Okbay et al.[1], including the UK Biobank. However, our project required non-overlapping GWAS samples for EA and schizophrenia (SZ). Therefore, all cohorts that were part of the most recent GWAS on SZ by the Psychiatric Genomics Consortium (PGC)[2] (deCODE, DIL, EGCUT, FTC, FVG, H2000, KORA, MGS, WTCCC58C) were excluded from the meta-analysis on EA, yielding a total sample size of n = 363,502. Our replication sample (the Göttingen Research Association for Schizophrenia – GRAS, see Supplementary Note section 6) was not part of the GWAS on EA.

2 GWAS on schizophrenia

The PGC shared GWAS summary statistics on SZ with us. The results are based on Ripke et al.[2], but excluded data from our replication sample (GRAS, see Supplementary Note section 6), yielding a total sample size of n = 34,409 cases and n = 45,670 controls.

3 Quality control

Data sources and quality control procedures for the GWAS on EA and SZ are described in Okbay et al.[1] and Ripke et al.[2], respectively. The original EA results file contained 12,299,530 genetic markers, compared to 17,221,718 in the SZ results file.

Before we proceeded with our proxy-phenotype analyses, we applied the following additional quality control steps:

1. To maximise statistical power, we excluded single nucleotide polymorphisms (SNPs) that were missing in large parts of the two samples. Specifically, we continued with SNPs that were available in at least 19 out of 50 cohorts in the SZ results[2,a] and in N > 200,000 in the EA meta-analysis[1]. This step excluded 3,778,914 and 6,369,138 genetic markers for EA and SZ, respectively.

2. We dropped SNPs that were not available in both GWAS results files. This step restricted our analyses to the set of available genetic markers that passed the quality-control filters in both the EA and the SZ GWAS results, leaving us with 8,403,560 autosomal SNPs.

3. We dropped 6 SNPs with non-standard alleles (i.e. not A, C, T, or G) and 2 SNPs with mismatched effective alleles. Furthermore, we dropped 163,272 SNPs in the first and the 99th percentile of the distribution of differences in minor allele frequency (MAF) in the two results files. This final step eliminated SNPs that were

[a] The actual N per SNP was not provided in the SZ GWAS summary statistics.

likely to be affected by coding errors, strand flips, or substantial differences in MAF in the EA and SZ samples.

The remaining 8,240,280 autosomal SNPs were used in the proxy-phenotype and prediction analyses described in Supplementary Note sections 4, 5, 7 and 8.

4 Selection of education-associated candidate SNPs

We conducted our proxy-phenotype analyses following a pre-registered analysis plan (https://osf.io/dnhfk/), using 8,240,280 autosomal SNPs that passed quality control (see section 3).

4.1 Setting the *P* value threshold for the proxy-based SNPs

Ideally, proxy-phenotype analyses should use a pre-specified *P* value threshold that maximises the expected number of true positive results in the look-up stage. The optimal threshold trades off between two opposing effects. On the one hand, a less stringent threshold yields a larger number of candidates that are forwarded to the second stage. A larger set of candidates is more likely to contain true positives. On the other hand, a larger number of candidates requires that a more stringent experiment-wide significance level needs to be applied in the second stage to adjust for multiple testing, which decreases power to pick out the true positives from among the set of candidates[3]. In principal, it is possible to calculate the optimal *P* value threshold based on the observed distribution of standardised effects sizes and standard errors in the GWAS results of the proxy, the genetic correlation between the two traits, their SNP-based heritabilities, and the sample size of the target phenotype. Given these parameters, one can infer the expected effect size of a genetic variant on the target from the results on the proxy phenotype, calculate the statistical power for the look-up, and approximate the number of expected true positive associations from the look-up at various *P* value thresholds[4].

In the current case, the value of such theoretical calculations is limited because the genetic correlation between EA and SZ does not reflect the actual genetic overlap between the two traits adequately. Specifically, Okbay et al.[1] report low, but significant, bivariate LD score regression estimates for the genetic correlation between EA and SZ of 0.08 ($P = 3.2\times10^{-4}$). In addition, the 74 genome-wide significant EA loci have only 51% sign concordance with the SZ results. This sign concordance pattern cannot be differentiated from what would be expected by chance for two traits that exhibit no genetic overlap. Yet, the same 74 EA-associated loci are strongly enriched for association with SZ (enrichment *P* value < 0.002), which strongly rejects the hypothesis of no genetic overlap between the two traits. This implies that standard formulas[4] to calculate the expected effect size of a genetic variant on SZ from the observed results on EA will be too conservative because they ignore the specific pattern of split sign concordance but strong enrichment for this pair of traits.

Instead of calculating a noisy theoretical optimum, we follow Rietveld et al.[3] and selected 10^{-5} as the default *P* value threshold prior to carrying out the proxy-phenotype analyses (https://osf.io/dnhfk/).

5 Look-up, sign concordance and enrichment

We used the GWAS results on EA and SZ (see Supplementary Note sections 1-2) that passed our quality control (Supplementary Note section 3) for the analyses described here.

5.1 Look-up

To select approximately independent SNPs from the EA GWAS results, we applied the clumping procedure in PLINK version 1.9[5,6] using $r^2 > 0.1$ and 1,000,000 kb as the clumping parameters and the 1000 Genomes phase 1 version 3 European reference panel[7] to estimate linkage disequilibrium (LD) among SNPs. This algorithm assigns the SNP with the smallest *P* value as the lead SNP in its "clump". All SNPs in the vicinity of 1,000,000 kb around the lead SNP that are correlated with it at $r^2 > 0.1$ are assigned to this clump. The next clump is formed around the SNP with the next smallest *P* value, consisting of SNPs that have not been already assigned to the first clump. This process is iterated until no SNPs remain with $P < 10^{-5}$, leading to 506 approximately independent EA-associated lead SNPs. 108 of the 506 EA-associated lead SNPs are genome-wide significant ($P < 5 \times 10^{-8}$).

We looked up the SZ GWAS results (Supplementary Note section 2) for these 506 EA-associated lead SNPs. Results for all 506 SNPs are reported in Supplementary Table 5.1. Figure 2a shows a Manhattan plot for the GWAS on EA highlighting SNPs that were also significantly associated with SZ (red crosses for $P_{SZ} < 0.05$, green crosses for $P_{SZ} < 0.05/506 = 9.88 \times 10^{-5}$). Figure 2b presents a Q-Q plot of the look-up. 132 SNPs are associated with SZ at nominal significance ($P < 0.05$) and 21 of these survive Bonferroni correction ($P_{SZ} < 9.88 \times 10^{-5}$).

In order to investigate the novelty of the findings, we extracted all the SNPs in LD with these 21 SNPs at $r^2 \geq 0.1$ with a maximum distance of 1,000 kb using the 1000 Genomes phase 1 European reference panel. Of these 21, 12 are in LD with loci previously reported by Ripke et al. and 2 are in the major histocompatibility complex (MHC) region on chromosome 6 and were therefore not separately reported in that study[2]. Three of the 21 loci that were not identified yet by Ripke et al.[2] were independently found in a recent meta-analysis of the PGC results with another large-scale cohort, yielding a total sample size of $n = 40{,}675$ SZ cases and $n = 64{,}643$ controls[8], strengthening the credibility of these loci. Furthermore, two are in LD with loci recently reported by Hellard et al.[9] who used a false discovery rate (FDR) method that is less conservative than our approach. The results of that latter study were also based on a much smaller GWAS sample for EA ($n = 95{,}427$) and partially overlapping samples between EA and SZ. Thus, our finding lends additional credibility to the suggestive associations reported in Hellard et al[9].

The remaining 2 SNPs we identified (tagged by rs7336518 on chr 13 and rs7522116 on chr 1) are credible candidate loci for SZ. Rs7522116 is in LD ($r^2 = 0.52$) with rs10218712, which reached suggestive significance in a previous study ($P = 4.0 \times 10^{-06}$)[10]. Similarly, rs7336518 is in LD ($r^2 = 0.67$) with rs11617058, which reached suggestive significance in the same study ($P = 1.0 \times 10^{-06}$)[10].

5.2 Bayesian credibility of results

We probed the credibility of our proxy-phenotype association results using a heuristic Bayesian calculation following Rietveld et al. (Supplementary Note pp. 13-15)[3]. We focus on the 21 EA-associated lead SNPs that are also associated with SZ after Bonferroni correction.

Bayes' Rule implies that the probability that an association is true given that we observe significance is given by

$$P(H_1 | t > t_{\alpha/2}) = \frac{P(t > t_{\alpha/2} | H_1) P(H_1)}{P(t > t_{\alpha/2} | H_1) P(H_1) + P(t > t_{\alpha/2} | H_0) P(H_0)}$$

$$= \frac{(power)(\pi)}{(power)(\pi) + (\alpha)(1 - \pi)}$$

"Power", as well as the significance test, are 2-sided, π is the prior belief that the SNP is truly associated, and α is the significance threshold used for testing (in our case, $\alpha = \frac{0.05}{506} = 9.88 \times 10^{-5}$).

To calculate power for each SNP, we computed the winner's curse corrected *odds* ratio using the procedure described in Rietveld et al. (Supplementary Note pp. 7-13)[11] for the α threshold of 9.88×10^{-5}. Because the actual sample size per SNP is not reported in the SZ GWAS summary statistics, we furthermore assumed that each SNP was available in the entire sample of 34,409 cases and 45,670 controls (i.e. the PGC results from Ripke et al.[2] excluding the GRAS data collection).

An important question is which prior beliefs are reasonable starting points for these Bayesian calculations. For an arbitrarily chosen SNP, the most conservative reasonable prior would assume that each truly associated SNP has the same effect size as the strongest effect size that was actually observed in the data. If one divides the SNP-based heritability of the trait by that effect size in R^2 units, one obtains a lower bound for the number of SNPs that can be assumed to be truly associated. To aid this line of thinking, we converted the winner's curse corrected *odds* ratios of our 21 SNPs into R^2 using

$$R^2 = \left(\frac{d}{\sqrt{d^2 + a}}\right)^2$$

where d is Cohen's *d*, which is calculated as

$$d = \ln(Odds)\frac{\sqrt{3}}{\pi}$$

and a is a correction factor that adjusts for the MAF of the SNP. This correction factor is calculated as

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2}$$

where $n_1 = N \times MAF$ and $n_2 = N \times (1 - MAF)$, see Borenstein et al.[12]

The largest effect size in R^2 that we observe in our results is rs4378243 with 0.044%. The SNP-based heritability of SZ is ≈21%[13]. Thus, if all causal SZ SNPs would have an effect of $R^2 = 0.044\%$, we would expect that ≈500 truly causal loci exist. The chance of finding any one of them by chance from a set of ≈500,000 independent loci in the human genome is

≈0.1%.[b] However, in reality most truly associated loci for SZ will surely have smaller effects than that. Thus, a prior belief of ≈0.1% is certainly too conservative.

Furthermore, the SNPs we investigate are *not* arbitrary but selected based on their association with another, genetically related cognitive trait (EA) in a very large, independent sample. Thus, a prior belief of 1% or 5% that these SNPs are also associated with SZ is probably more reasonable. As an upper bound, we assume that 10% of all loci are causal. Thus, the chance to pick any one of them by chance would be 10%.

Table 1 displays the winner's curse corrected effect size of the 21 EA-associated lead SNPs that are also associated with SZ after Bonferroni correction. It also shows the posterior probability that these SNPs are truly associated with SZ given our results for prior beliefs ranging from 0.1%, 1%, 5%, to 10%. Thirteen of these SNPs have posterior probabilities of being true positives of >50% for even the most conservative prior. For a more realistic prior belief of 5%, all 21 SNPs are likely or almost certain to be true positives.

5.3 Sign concordance

We compared the signs of the beta coefficients of the 506 EA lead SNPs ($P_{EA} < 10^{-5}$) with the beta coefficients for SZ. If the signs were aligned, we assigned a "1" to the SNP and "0" otherwise. By chance, sign concordance is expected to be 50%. We tested if the observed sign concordance is different from 50% using the binomial probability test[14]. 263 of the 506 SNPs have the same sign (52%, P = 0.40, 2-sided).

Sign concordance is 58% (P = 0.10, 2-sided) in the set of 132 EA lead SNPs that are also nominally significant for SZ ($P_{EA} < 1 \times 10^{-5}$ and $P_{SZ} < 0.05$).

Finally, for the 21 SNPs that passed Bonferroni correction for SZ ($P_{EA} < 1 \times 10^{-5}$ and $P_{SZ} < 9.88 \times 10^{-5}$), sign concordance is 62% (P = 0.38, 2-sided).

5.4 Enrichment

Because EA and SZ are highly polygenic, we tested for enrichment by taking the actual distribution of P values in the GWAS result files into account.

5.4.1 Raw enrichment factor (not corrected for LD score of SNPs)

Due to the polygenic architecture of both traits, it is expected to find some EA-associated SNPs that are also associated with SZ just by chance even if both traits are genetically independent. Under this null hypothesis, the expected number of EA-associated lead SNPs that are also significantly associated with SZ is

$$E_{H_0}\left[N_{S,EA\rightarrow SZ}\right] = N_{T,EA} \times \tau_{P_{EA}} \times \tau_{P_{SZ}}$$

where $N_{T,EA}$ is the total number of independent lead SNPs in the EA GWAS results, and $\tau_{P_{EA}}$ and $\tau_{P_{SZ}}$ are the shares of SNPs in $N_{T,EA}$ that have P values for EA and SZ below a certain threshold, respectively.

[b] It is typically assumed that GWAS data for European populations contain ≈1,000,000 independent loci. However, the quality-control procedures for GWAS summary statistics in studies like ours decreases the number of independent loci to <1,000,000[41,62]. In fact, clumping the post-QC GWAS results for SZ without a P value threshold, an R^2_{LD}<0.1, and a LD-window of 1,000,000 kb with the 1000 Genomes phase 1 version 3 European reference panel[42] leads to only 223,065 independent loci. Thus, assuming 500,000 independent loci in these calculations is conservative.

We define the raw enrichment factor as

$$N_{S,EA\to SZ}/E\left[N_{S,EA\to SZ}\right]$$

where $N_{S,EA\to SZ}$ is the observed independent number of SNPs that pass both the *P* value thresholds P_{EA} and P_{SZ}.

We obtained $N_{T,EA}$ by applying the clumping procedure described in Supplementary Note section 5.1 without a *P* value threshold for EA, leading to 222,289 independent EA lead SNPs in our merged results file (Supplementary Note section 3). For $P_{EA} < 10^{-5}$, we found 506 SNP ($\tau_{P_{EA}} = \frac{506}{222{,}289} = 0.2276\%$).

The Bonferroni threshold for testing 506 independent hypothesis is $P_{SZ} < \frac{0.05}{506} = 9.88 \times 10^{-5}$. There are 341 independent SNPs in the SZ results that pass this threshold, thus $\tau_{P_{SZ}} = \frac{341}{222{,}289} = 0.1534\%$. Therefore, we expect $\left[N_{S,EA\to SZ}\right] = 222{,}289 \times 0.2276\% \times 0.1534\% = 0.776$ (i.e. less than one) SNP to be jointly associated with both traits under the hull hypothesis of no genetic overlap. At these *P* value thresholds, we actually observe $N_{S,EA\to SZ} = 21$ SNPs, implying a raw enrichment factor of $\frac{21}{0.776} = 27$.

For $P_{SZ} < 0.05$, we found 17,935 SNP ($\tau_{P_{SZ}} = \frac{17{,}935}{222{,}289} = 8.068\%$). Thus, $\left[N_{S,EA\to SZ}\right] = 222{,}289 \times 0.2276\% \times 8.068\% = 41$. At this more liberal *P* value threshold, we actually observe $N_{S,EA\to SZ} = 132$ SNPs, implying a raw enrichment factor of $\frac{132}{41} = 3.23$.

5.4.2 Raw enrichment P value (not corrected for LD score of SNPs)

Following Okbay et al.[4], we performed a non-parametric test of joint enrichment that probes whether the EA lead SNPs are more strongly associated with SZ than randomly chosen sets of SNPs with MAF within one percentage point of the lead SNP. To perform our test, we randomly drew 10 matched SNPs for each of the 506 EA lead SNPs with $P_{EA} < 10^{-5}$.

We then ranked the 506×10 randomly matched SNPs and the original 506 lead EA SNPs by *P* value and conducted a Mann-Whitney test[15] of the null hypothesis that the *P* value distribution of the 506 EA lead SNPs are drawn from the same distribution as the 506×10 randomly matched SNPs. We reject the null hypothesis with $P = 6.872 \times 10^{-10}$ ($Z = 6.169$, 2-sided). As a negative control test, we also calculated the raw enrichment *P* value of the first randomly drawn, MAF-matched set of SNPs against the remaining 9 sets, yielding *P* = 0.17.

Repeating this raw enrichment test for the subset of 21 EA-associated SNPs that remained significantly associated with SZ after Bonferroni correction (threshold $P_{SZ} < \frac{0.05}{506} = 9.88 \times 10^{-5}$) yields $P = 5.44 \times 10^{-14}$ ($Z = 7.521$, 2-sided). The negative control test based on the raw enrichment *P* value of the first randomly drawn, MAF-matched set of SNPs against the remaining 9 sets yields *P* = 0.34.

5.4.3 LD-aware enrichment test

The "raw" enrichment reported in Supplementary Note section 5.4.1 could in principle be due to the LD-structure of the EA lead SNPs that we tested. Specifically, if these EA lead SNPs have stronger LD with other SNPs in the human genome than expected by chance, this could cause the observed enrichment of this set of SNPs on SZ and other traits because higher LD

increases the chance these SNPs would "tag" causal SNPs that they are correlated with. To test whether the observed enrichment is due to LD, we developed an LD-aware enrichment test. Furthermore, we used this LD-aware enrichment test to probe if the observed enrichment of our EA lead SNPs can also be observed for other traits.

We investigated the SZ GWAS results described in Supplementary Note section 2 and 21 additional traits for which GWAS summary statistics were available in the public domain (Supplementary Table 5.2). Some of the traits were chosen because they are phenotypically related to SZ (bipolar disorder (BIP), neuroticism, depressive symptoms, major depressive disorder, autism, and childhood intelligence (IQ)), while others were less obviously related to SZ (e.g. intracranial volume, cigarettes per day) or to the brain (e.g. age at menarche, inflammatory bowel disease). Finally, we included five traits as negative controls (height, birth weight, birth length, fasting (pro)insulin).

We calculated the LD score regression intercept and slope of the traits using LDSC[16]. For SNP *i* in trait *j*, the expected chi-square statistic can be calculated as

$$E[{Z_{ij}}^2] = ({N_j} \times {h^2}_j \times LDscore_i/M) + (1 + Na)_j$$

where *N* is the sample size of the target trait *j*, h^2 is the heritability of trait *j*, $LDscore_i = \sum_{k=1}^{M} r_{ik}^2$ for SNP *i* is calculated using HapMap3 SNPs from European-ancestry, *M* is the number of SNPs included in the calculation of the LD score (*n* = 1,173,569 SNPs), r_{jk}^2 is the squared correlation between SNPs *j* and *k* in the HapMap3 reference panel, and *1 + Na* is the LD score regression intercept for trait *j*. We used precomputed LD scores available from the LDSC software[16].

As recommended by Bulik-Sullivan et al.[16], we restricted our analysis to HapMap3 SNPs (using the --merge-alleles flag) because these seem to be well-imputed in most studies. Out of 132 SNPs with $P_{EA} < 1 \times 10^{-5}$ and $P_{SZ} < 0.05$, only 30 SNPs are directly present in HapMap3 SNP list. Therefore, we extracted proxy SNPs with $r^2 > 0.8$ and a maximum distance of 500 kb to our missing EA lead SNPs and chose the one with the highest r^2 as a proxy. After this step, we could include 105 (out of 132) SNPs in our analyses. For each of these 105 SNPs, we observed the *Z*-statistics in the publically available GWAS results of the traits. *Z*-statistics were converted into Chi^2 statistics by squaring them. The LD score corrected enrichment per SNP for each trait is the ratio of the observed to the expected Chi^2. The results are shown in Figure 3 (and in Supplementary Table 5.2). To test whether a particular realization is significantly larger than expected (and thus the ratio $Chi^2_{observed}/Chi^2_{expected}$ is significantly greater than one), we test each particular observed Chi^2 against a non-central Chi^2 distribution with $k = 1 + Na$ degrees of freedom and a non-centrality parameter of $N \times h^2 \times LDscore_i/M$. The intuition behind using the LD score regression slope as the non-centrality parameter is as follows: True genetic signal, not caused by stratification, covaries with LD[16]. True genetic signal also results in non-central Chi^2 statistics (i.e. $\beta \neq 0$ leads to non-centrality in $(\frac{\beta}{se})^2$). Therefore, under a polygenic model, SNPs with a high LD score are expected to tag several true effects, raising our expectation of their Chi^2 statistic. Here we account for the LD score specific expectation of effect size. We show the selected SNPs influence our traits of interest over and above what is expected based on LD and, importantly, our results suggest that this is not the case for several other phenotypes, revealing a certain amount of specificity in the enriched SNPs.

Furthermore, since the SNPs considered for enrichment are independent, their Chi^2 and non-centrality parameters are additive. This additivity allows us to formulate an expected distribution of the sum of $Chi^2_{observed}$, based on the sum of non-centrality and the sum of

degrees of freedom for the 105 SNPs. Against this expected distribution we can test the observed sum of Chi^2 statistics for all SNPs. This test provides us with a *P* value for the LD-aware enrichment test. The *P* value reflects excess of the enrichment for the set of SNPs beyond what is expected if these SNPs are part of the infinitesimal genetic contribution to the trait in question.

Our LD-aware enrichment test has two limitations. First, LD score regression assumes that allele frequency (AF) does not correlate with effect size, an assumption which has been empirically shown to be violated for low-frequency alleles[17]. Second, our test assumes the absence of selection on the trait. Variation in AF and the degree of negative selection could explain excess signal in low LD SNPs[18]. However, our raw enrichment *P* value (see Supplementary Note section 5.4.2) is robust to this because it takes the AF of the candidate SNPs explicitly into account.

Supplementary Table 5.2 and Figure 3 summarise the results. We find that the enrichment of EA-associated SNPs for SZ cannot be explained by the LD scores of these SNPs: The set of 105 SNPs is jointly associated with SZ after controlling for their LD scores ($P < 4 \times 10^{-14}$) and 15 of these SNPs are individually associated with SZ after Bonferroni correction. We also observe LD score corrected enrichment of these SNPs with several other phenotypes, most noticeably with BIP (joint $P = 1.1 \times 10^{-16}$). Four out of 93 tested SNPs are significantly associated with BIP after Bonferroni correction, including one of the SNPs that our proxy-phenotype analysis isolated as a novel candidate locus for SZ (rs9575628, a proxy for rs7336518, see Supplementary Table 5.3). We also observe weaker, but still significant joint LD score corrected enrichment of this set of SNPs for inflammatory bowel disease, neuroticism, age at menarche, and childhood IQ. Note that several brain-phenotypes do not show significant enrichment, including depressive symptoms, major depressive disorder, ADHD, and Alzheimer's disease. This implies that the set of SNPs we are testing exhibits some phenotype-specificity is not simply involved in all brain-related outcomes. Also note that none of the negative control phenotypes we included shows significant LD score corrected enrichment (height, birth weight, birth length, fasting (pro)insulin).

5.5 Prediction of future genome-wide significant loci for schizophrenia

As reported above (Supplementary Note section 5.1), three of the SNPs that our proxy-phenotype approach identified after Bonferroni correction have been reported as novel, genome-wide significant loci for SZ in an effort[8] that was ongoing parallel to ours. Overall, 50 novel loci for SZ were reported in that study. This provides us with the opportunity to ask if our proxy-phenotype approach using GWAS results from EA was able to predict "future" GWAS findings for SZ.

We are using a simple proportions test for this purpose, which compares the ratio of novel SNPs included in our list of 132 loci that are jointly associated with EA and SZ ($P_{EA} < 10^{-5}$ and $P_{SZ} < 0.05$) with the ratio observed in all remaining approximately independent loci with $P_{SZ} < 0.05$.

To identify LD partners and to clump our GWAS results, we used a threshold of $r^2 > 0.1$ and a 1,000,000 kb window in the 1000 Genomes phase 1 version 3 European reference panel. Our SZ summary statistics contained 51,721 approximately independent SNPs with $P_{SZ} < 0.05$. We identified 21,430 SNPs in LD with the 50 novel SNPs[8] and 54,425 SNPs in LD with the 128 genome-wide significant loci that were previously reported.[2] We removed SNPs in LD with the previously GWAS hits from our analyses because those SNPs could (by definition) not be identified as novel. The remaining set of 51,528 approximately independent

SNPs with $P_{SZ} < 0.05$ in our SZ GWAS results contained one proxy for each of the 50 novel SNPs[8]. 110 SNPs with $P_{SZ} < 0.05$ also exhibited $P_{EA} < 10^{-5}$ in the independent EA GWAS sample. Of those 110 SNPs, six were identified as novel SZ loci in the most recent GWAS dataset expansion[8]. Using Fisher's exact test, we rejected the null hypothesis that the proportion of novel SNPs is equal in the two sets ($P = 4.1 \times 10^{-8}$, 2-sided). Furthermore, as a robustness check, we performed the analysis again by excluding the SNPs with MAF ≤ 0.1 and found similar results ($P = 1.2 \times 10^{-6}$). Thus, we conclude that conditioning GWAS results on SZ with independent GWAS evidence on EA significantly outperforms pure chance in predicting GWAS results on SZ from even larger samples.

6 The GRAS data collection

All parts of GRAS data collection comply with the Helsinki Declaration and were approved by the ethical committee of the Georg-August-University of Göttingen (master committee) as well as by the respective local regulatory/ethical committees of all collaborating centres.

6.1 Subjects

GRAS schizophrenia and schizoaffective patients

The GRAS data collection has been established over the last 10 years and consists of >1,200 deep phenotyped patients, diagnosed with SZ or schizoaffective disorder (according to *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition Text Revision [DSM-IV-TR]) recruited across 23 collaborating centres across Germany[19,20]. All patients and/or their authorised legal representatives gave written informed consent. For the present study 1,067 patients were included, 67.1% were male (n = 716) and 32.9% female (n = 351). The average age was 39.46 ±12.58 years (range 17-79).

GRAS healthy controls

Healthy controls, who gave written informed consent, were voluntary blood donors, recruited by the Department of Transfusion Medicine of the George-August-University of Göttingen (Germany) according to national guidelines for blood donation. As such, they widely fulfil health criteria, ensured by a broad predonation screening process containing standardised questionnaires, interviews, haemoglobin, blood pressure, pulse, and body temperature determinations[19]. Participation as healthy controls for the GRAS data collection was anonymous, with information restricted to age, gender, blood donor health state and ethnicity. For the present study 1,169 subjects were included, 62.2% were male (n = 727) and 37.8% female (n = 442). The average age was 37.44 ±13.27 years (range 18-69).

6.2 Genotyping

Genotyping of the GRAS patients and control sample was done with a semi-custom Axiom myDesign genotyping array (Affymetrix, Santa Clara, CA, USA), based on a CEU (Caucasian residents of European ancestry from Utah, USA) marker backbone including 518,722 SNPs, and a custom marker set including 102,537 SNPs. The array was designed using the Axiom Design centre, applying diverse selection criteria[21]. Genotyping was done by Affymetrix on a GeneTitan platform. Several quality control steps were used (SNP call rate >97%, Fisher's linear discriminant >3.6, heterozygous cluster strength off set > -0.1, and

homozygote ratio off set > -0.9). These steps were completed with use of either genotyping console software (Affymetrix) or R. In a subsequent step, markers in X, Y, and mitochondrial chromosomes and those with Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$ (GRAS healthy controls) or $P < 1 \times 10^{-10}$ (GRAS patients) were removed, leaving 589,921 SNPs available for analyses.

6.3 Imputation and estimation of genetic principal components

The full details of genotype imputation and the estimation of the genetic principal components in the GRAS data collection are described elsewhere[2]. Briefly, imputation was done with the prephasing or imputation approach implemented in IMPUTE2 and SHAPEIT (chunk size 3 Mb)[22,23]. A version of the phase 1 integrated variant set release (v3) from the full 1000 Genomes Project dataset (March 2012) that is limited to variants with more than one minor allele copy ("macGT1"; Aug 26, 2012) was used as imputation reference dataset (INFO value > 0.1 and MAF >0.005). Ten principal components of the genetic data in the GRAS sample were obtained using the standard PGC protocol[2].

6.4 Phenotyping procedures

Patients from the GRAS data collection were examined by the GRAS team of travelling investigators after giving written informed consent, own and/or authorised legal representatives. The GRAS team of travelling investigators consisted of 1 trained psychiatrist and neurologist, 3 psychologists and 4 medical doctors/last year medical students. All investigators had continuous training and calibration sessions to ensure the highest possible agreement on diagnoses and other judgments as well as a low interrater variability regarding the instruments applied. A full description of the GRAS data collection standard operating procedures is provided elsewhere[20].

Deep clinical phenotyping data were available for all the GRAS patients. For the purpose of the present study the following phenotypes for SZ were selected: i) **age at prodrome** which precedes SZ onset and is characterized by cognitive decline, social withdrawal, and depression; ii) **age at disease onset** defined as onset of first psychotic episode; iii) **premorbid IQ** using the MWT-B (Mehrfachwahl-Wortschatz-Intelligenztest-B) which estimates the intellectual functioning of a person prior to known or suspected onset of disease[24]; iv) positive and negative symptoms using the **Positive and Negative Syndrome Scale (PANSS)**[25]. Each domain ranges from 7 to 49 with higher scores indicating severe symptoms; v) the **Global Assessment of Functioning (GAF)** which subjectively rates the social, occupational, and psychological functioning of an individual, e.g., how well one is meeting various problems-in-living, on a continuum ranging from 1 to 100[26]. Poorer functioning is indicated by lower GAF scores; vi) the **Clinical Global Impression of Severity (CGI-S)** scale which rates illness severity from 1 to 7 with higher scores indicating more severe illness.[27]; vii) **years of education** was measured based on the highest degree obtained, converted into US-schooling year equivalents based on the 1997 International Standard Classification of Education (ISCED) of the United Nations[28]. Education was assessed retrospectively at the time of patient recruitment for the GRAS data collection. At this time, 75% of the GRAS patients were older than 29 years.

An important feature of SZ patients that may influence their everyday functioning and performance, and result in a considerable number of side effects, is their antipsychotic medication. Medication was assessed as the dose of present antipsychotic medication of each patient at the moment of the interview, expressed as chlorpromazine equivalents[29].

The sociodemographic and clinical characteristics of the GRAS data collection are reported in Supplementary Table 6.1.

7 Replication in the GRAS data collection

We showed in our pre-registered analysis plan that our replication sample (GRAS) is not large enough to replicate individual SNPs (https://osf.io/dnhfk/). Instead, we decided at the outset to attempt replication of the proxy-phenotype analysis results using a polygenic score (PGS) that consists of the >80 most strongly associated, independent SNPs. The set that best meets this criterion are the 132 independent EA lead-SNPs that are also nominally associated with SZ ($P_{SZ} < 0.05$), see Supplementary Note section 5. PGS for this set of 132 candidate SNPs were constructed using either the β coefficient estimates of the EA or the SZ GWAS meta-analysis, resulting in two different scores (named *EA_132* and *SZ_132* in Supplementary Tables 7.1.a, b and c, 7.2, 8.2.a and b).

In addition, we also constructed PGS for EA, SZ, BIP, and neuroticism in the GRAS data collection using all available SNPs as control variables for multivariate prediction analyses (named *EA_all*, *SZ_all*, *BIP_all*, and *Neuro_all* in Supplementary Tables 7.1.a, b and c, 7.2, 8.2.a and b, 8.3, 8.4.a and b). Technical details are described below.

7.1 Polygenic score calculations

PGS were calculated using PLINK version 1.9[5,6]. We calculated eight different scores, which are described below.

7.1.1 Schizophrenia scores

We received the GWAS summary statistics for SZ from the PGC excluding the data from our replication sample (GRAS). We constructed a PGS using the 132 EA lead-SNPs ($P_{EA} < 10^{-5}$) that are also nominally associated with SZ ($P_{SZ} < 0.05$). This score (*SZ_132*) is used for replication of the proxy-phenotype analyses described in Supplementary Note section 5.

Furthermore, we constructed a PGS using all 8,240,280 SNPs that survived quality control (*SZ_all*, see Supplementary Note section 3). Next, we applied the clumping procedure using r2 > 0.1 and 1,000 kb as the clumping parameters and the 1000 Genomes phase 1 version 3 European reference panel to estimate LD among SNPs, eventually leaving a set of 349,357 SNPs ready for profile scoring.

For secondary analyses on the prediction of SZ symptoms (Supplementary Note section 8.3), we constructed two PGS using the 4,147,926 SNPs and 4,092,354 SNPs that have concordant (+ and +; or – and –) or discordant signs (+ and –; or – and +) for EA and SZ, respectively. Clumping resulted in 260,441 and 261,062 independent SNPs with concordant or discordant, respectively. We used these approximately independent SNPs for profile scoring and call the resulting PGS *Concordant* and *Discordant*.

These four scores were calculated using the –score function in PLINK using the natural log of the *odds* ratio of the SNPs for SZ.

7.1.2 Educational attainment scores

Beta coefficients for the EA GWAS described in Supplementary Note section 1 were approximated using $\widehat{\beta_j} = \frac{z_j}{\sqrt{N_j * 2 * MAF_j * (1-MAF_j)}}$, see Rietveld et al. (Supplementary Note pp. 4-6)[11] for the derivation. Using these betas, we constructed a PGS using the 132 EA lead-SNPs ($P_{EA} < 1 \times 10^{-5}$) that are also nominally associated with SZ ($P_{SZ} < 0.05$). The resulting score is called *EA_132*.

Furthermore, we constructed a PGS using all 8,240,280 SNPs that survived quality control (see Supplementary Note section 3). Next, we applied the clumping procedure using $r^2 > 0.1$ and 1,000 kb as the clumping parameters and the 1000 Genomes phase 1 version 3 European reference panel to estimate linkage disequilibrium among SNPs, eventually leaving a set of 348,429 SNPs ready for profile scoring. The resulting score is called *EA_all*.

7.1.3 Bipolar disorder score

We obtained GWAS summary statistics on BIP from the PGC[30]. We used the LD-pruned GWAS summary from PGC ("pgc.bip.clump.2012-04.txt") with a set of 108,834 LD-pruned SNPs ready for profile scoring. PGS for the GRAS data collection were calculated by the application of the *–score function* in PLINK using the natural log of the *odds* ratio. The resulting score is called *BIP_all*.

7.1.4 Neuroticism score

We obtained GWAS summary statistics on Neuroticism from the SSGAC. The results are based on the analyses reported in Okbay et al.[4] containing 6,524,432 variants. We applied the clumping procedure using $r^2 > 0.1$ and 1,000 kb as the clumping parameters and the 1000 Genomes phase 1 version 3 European reference panel to estimate LD among SNPs, eventually leaving a set of 232,483 SNPs ready for profile scoring. PGS for the GRAS data collection were calculated by the application of *–score function* in PLINK using the Neuroticism beta values. The resulting score is called *Neuro_all*.
Note that our replication sample (GRAS) was not included in the GWAS summary statistics of any of these traits.

7.2 Polygenic score correlations

We calculated Pearson correlations between all PGS that we constructed in the GRAS data collection (*SZ_all*, *SZ_132*, *EA_all*, *EA_132*, *Concordant*, *Discordant*, *BIP_all*, and *Neuro_all*). Results for SZ patients and healthy controls together are reported in Supplementary Table 7.1.a. *SZ_all* was positively correlated with *SZ_132* ($r = 0.188$; $P < 0.0001$) and *EA_all* positively correlated with *EA_132* ($r = 0.148$, $P < 0.0001$). We found a high correlation between the *SZ_all* score and *Concordant* ($r = 0.871$, $P < 0.0001$) and *Discordant* ($r = 0.881$, $P < 0.0001$) scores. Furthermore, we found a moderate correlation between *EA_all* and *Concordant* ($r = 0.256$, $P < 0.0001$) as well as between *EA_all* and *Discordant* ($r = -0.377$, $P < 0.0001$) scores. *Concordant* and *Discordant* scores showed a highly positive correlation ($r = 0.627$, $P < 0.0001$). We found very similar results among the SZ cases (Supplementary Table 7.1.b) and healthy controls when we analyzed them separately from each other (Supplementary Table 7.1.c). These results were used to inform the correct multiple regression model specification for the analyses presented in section 8.3.

7.3 Predicting case-control status using PGS

To investigate if our proxy-phenotype results help to predict SZ case-control status, we estimated a linear probability model (LPM) in Stata[31,32]. Genetic outliers (n = 13) based on self-reported non-European ancestry were excluded from all prediction analyses. Following the PGC protocol described in Ripke et al.[2], we included 10 principal components as covariates (but no other variables). The proportion of variance explained (*Adjusted* R^2) was computed by the comparison of a full model (covariates + PGS) to a reduced model (covariates only).

Results of the predictions with different models are summarised in Table 2. As effect sizes, we are reporting standardised regression coefficients. *SZ_132* scores significantly predict the case-control status in our replication sample ($P = 5.4 \times 10^{-0}$) (Table 2 Model 1) and remain significant even if we include *SZ_all* score as a control variable ($P = 1.7 \times 10^{-0}$) (Table 2 Model 3). The *EA_132*, *EA_all* and *Neuro_all* scores did not predict case-control status in our replication sample. The *BIP_all* score significantly predicts case-control status (Table 2 Model 7), which is expected given the genetic correlation between BIP and SZ[33,34]. Interestingly, the *SZ_132* score still significantly predicts case-control status when all other scores are included as control variables ($P = 3.4 \times 10^{-0}$) (Table 2 Model 9), highlighting the importance of these 132 SNPs. Furthermore, as a robustness check we excluded the 132 SNPs from the construction of the *SZ_all* and *EA_all* scores. The prediction results did not change except a slight decrease of variance explained with *SZ_all* scores, which is expected given that 132 SNPs were excluded from the construction of the score (Supplementary Table 7.2).

As an additional robustness check, we also ran a logistic regression with standardised PGS to predict SZ case-control status using the same explanatory variables as in Models 1-9 in Table 2. In all models, the *P* values of the coefficients were very similar to the ones obtained by LPM. In the equivalent of Model 9, we find that increasing the PGS from its mean by one standard deviation increases the probability of being an SZ case in the GRAS sample by 4.3% for the *SZ_132* score ($P = 3.1 \times 10^{-0}$), 15.2% for the *SZ_all* score ($P = 3.3 \times 10^{-0=}$), and 7.1% for the *BIP_all* score ($P = 4.1 \times 10^{-0=}$), respectively.

8 Polygenic prediction of schizophrenia symptoms

8.1 Phenotypic correlations

Phenotypic correlations between years of education and different phenotypes of SZ in the GRAS data collection (see Supplementary Note section 6) were calculated using Pearson's correlation. Target SZ phenotypes included were age at prodrome, age at disease onset, premorbid IQ, GAF, CGI-S, and positive and negative symptoms of the PANSS.

Results from the phenotypic correlations between the target phenotypes are reported in Supplementary Table 8.1 and Supplementary Figure 1. In summary, years of education was significantly correlated (all *P* values <0.0003) with all the quantitative traits of interest for the present study measuring the psychopathology and level of functioning of the patients (-0.212 < r < 0.435). The correlation was positive for age at prodrome and age at disease onset, indicating that the earlier the disease started the lower the level of education achieved. Years of education was also positively correlated with premorbid IQ and GAF and negatively

correlated with CGI-S and PANSS positive and negative scores, indicating that higher levels of education were associated with less severe disease outcomes. Our results are in line with previous studies suggesting that less educated SZ patients are at higher risk of a poorer course of the disease[35].

Moreover, we found a strong positive correlation between age at prodrome and age at disease onset (r = 0.918; P < 0.0001). Premorbid IQ was positively correlated with GAF (r = 0.200; P < 0.0001) and negatively correlated with CGI-S, PANSS positive and negative (-0.277 < r < -0.110; all P values < 0.0009). Measures of disease severity (CGI-S, PANSS positive, PANSS negative) were positively correlated with each other (0.478 < r < 0.638; all P values < 0.0001) and negatively correlated with an assessment of global functioning (GAF, -.579 < r < -0.824; all P values <0.0001).

8.2 Based on the 132 EA lead-SNPs

To investigate if our proxy phenotype results can predict specific SZ features, we predicted eight quantitative outcomes among the SZ cases in the GRAS data collection: years of education, age at prodrome, age at disease onset, premorbid IQ, GAF, CGI-S, PANSS positive and PANSS negative scores. These phenotypes are described in Supplementary Note section 6.

SZ and BIP, although being classified as two different disorders, share several symptoms. The PANSS is a clinical instrument principally developed for use in SZ to identify the presence and severity of psychopathology symptoms. However, BIP patients also present some symptoms measured by this scale manifesting the phenotypic overlap between both diseases[36,37]. Thus, we used the sum score of the positive symptoms (delusions, conceptual disorganization, hallucinations, hyperactivity, grandiosity, suspiciousness/persecution, hostility) and the sum score of the negative symptoms (blunted affect, emotional withdrawal, poor rapport, passive/apathetic social withdrawal, difficulty in abstract thinking, lack of spontaneity and flow of conversation, stereotyped thinking) included in PANSS to test if the genetic associations we identified using our proxy-phenotype approach predict symptoms that SZ and BIP share.

For the prediction of each phenotype a linear regression model was used including the PGS described in Supplementary Note section 7 (*SZ_132, SZ_all, EA_132, EA_all, BIP_all and Neuro_all*). Each regression included 10 principal components as covariates. The regressions for the prediction of years of education and premorbid IQ also included the age of onset as a covariate. Furthermore, the regressions for the prediction of GAF, CGI-S and the PANSS scores also controlled for medication because medication significantly affects these outcomes. We calculated the marginal R^2 of each PGS by squaring its standardised beta coefficient. Results of the predictions have been summarised in Supplementary Table 8.2.a. We found that both the *EA_all* (*stand. beta* = 0.17, $P = 8 \times 10^{-0}$) and the *EA_132* score (*stand. beta* = 0.08, $P = 9 \times 10^{-0}$) were associated with years of education in the GRAS sample of SZ patients. However, the *EA_132* score did not survive correction for multiple testing (8 phenotypes and 6 PGS = 48 tests; thus the Bonferroni-adjusted P value is $\frac{0.05}{48} = 1.04 \times 10^{-0}$. However, we note that neither the phenotypes nor the PGS are strictly independent. Therefore, the Bonferroni correction is likely to be too conservative to obtain a family-wide error rate of 0.05). Since age of disease onset, disease severity and progress of the disease can have an effect on the level of education achieved by a SZ patient, the years of education assessed in the GRAS data collection is not as solid as the measure used in the most recent EA GWAS[1]. The potential measurement error of EA in the GRAS data collection

may therefore lead to an underestimation of the predicted association. The *EA_all* score was also associated with premorbid IQ (*stand. beta* = 0.14, $P = 1.8 \times 10^{-0}$). None of the PGS we constructed were associated with any of the other phenotypes tested (age at prodrome, age at disease onset, GAF, CGI-S, PANSS positive and negative scores) at $P < 0.03$.

As a robustness check, we excluded the 132 EA lead-SNPs from the proxy-phenotype analyses from the construction of the *EA_all* and *SZ_all* scores to correct for double-counting of these SNPs [*EA_all (without 132)* and *SZ_all (without 132)*]. All prediction models have been analysed again with these PGSs, yielding virtually identical results as our main model specification (Supplementary Table 8.2.b).

8.3 Based on the sign concordance between EA and SZ GWAS results

If heterogeneity in the genetic architecture of SZ subtypes is causing the observed enrichment of EA-associated loci with SZ, the sign concordance pattern of SNPs with both traits may contain relevant information that is pertinent to specific SZ symptoms. We tested this by constructing PGS that take the sign concordance of SNPs for both traits into account. Specifically, we took SNPs and SZ GWAS results that were used to construct the *SZ_all* score and sorted the SNPs into two sets based on their sign concordance with EA. The resulting two sets of SNPs were used to construct the *Concordant* and *Discordant* scores (for details, see Supplementary Note section 7.1.1).

A linear regression model was used for the prediction of each phenotype including the PGS described in Supplementary Note section 7 (*Concordant*, *Discordant*, *EA_all*, *BIP_all* and *Neuro_all*). Due to the high correlation between the *SZ_all* score with the *Concordant* (r = 0.858, $P < 0.0001$) and *Discordant* (r = 0.873, $P < 0.0001$) scores in the GRAS sample of patients (Supplementary Table 7.1.b), we excluded the *SZ_all* from these prediction models to avoid multicollinearity. Each regression included the first 10 genetic principal components as covariates. The regressions for the prediction of years of education and premorbid IQ also included the age of onset as a covariate. Furthermore, the regressions for the prediction of GAF, CGI-S and the PANSS scores also controlled for medication because medication significantly affects these outcomes.

Results of the predictions are summarised in Table 3. As expected, the *EA_all* score was associated with years of education and premorbid IQ accounting for 4.2% ($P = 2.6 \times 10^{-0}$) and 3% ($P = 2.3 \times 10^{-0}$) of the variance, respectively. Interestingly, with this new model we could also predict SZ symptoms such as GAF and PANSS (Table 3). For example, all five PGS jointly achieve an $\Delta R^2 = 1.2\%$ for PANSS negative and $\Delta R^2 = 1.4\%$ for GAF. Conditional on the effects of the *Concordant* and *Discordant* scores, the *EA_all* score is now associated with less severe disease outcomes, consistent with the observed phenotypic correlations. And conditional on the *EA_all* score, the *Concordant* score is now associated with more severe positive and negative symptoms as measured by the PANSS scale, worse global functioning measured by GAF, and higher illness severity measured by the CGI-S. Although a Bonferroni correction for multiple testing is too conservative for our models given that PGS and SZ phenotypes are not independent, 5 of the PGS survive Bonferroni correction ($P = 0.05/(5 \times 8) = 1.25 \times 10^{-0}$, see Table 3).

Results from linear regression models excluding the *EA_all* scores (Supplementary Table 8.3) suggest that the association between *EA_all* and disease outcomes is conditional on the effects of *Concordant* and *Discordant* scores. In the same way, the associations between *Concordant* and *Discordant* and disease outcomes are conditional on the effects of *EA_all* scores.

Although the correlation between *EA_all* and the *Concordant* (r = 0.256) and *Discordant* (r = -0.377) scores was only moderate (Supplementary Table 7.1.b), we checked for possible multicollinearity in all our prediction models. The variance inflation factor was less than 3 for all the models, suggesting that multicollinearity is not a concern for the prediction results reported in Table 3[38].

Since the GRAS data collection includes SZ and schizoaffective disorder (SD) patients, we repeated analyses described above excluding patients that were diagnosed with SD (n = 198). We found that the 95% confidence intervals of the estimated regression coefficients of both model specifications overlapped in all cases, implying that the genetic heterogeneity in SZ that we identify is not only due to SD (Supplementary Table 8.4.a and b).

9 Controlling for genetic overlap between schizophrenia and bipolar disorder

9.1 GWIS schizophrenia – bipolar disorder

One possible reason for the observed genetic overlap between EA and SZ is that both phenotypes could be jointly genetically correlated with other outcomes. In fact, Nieuwboer et al.[39] suggest that the genetic correlation between EA and SZ is probably induced by the genetic correlation between SZ and BIP as well as the genetic correlation between BIP and EA. If that is indeed the case, the EA-associated lead SNPs should not show enrichment for association in "unique" SZ GWAS results that are "purged" of the genetic correlation between SZ and BIP.

To test this hypothesis, we estimated genome-wide inferred statistics (GWIS)[39] to obtain SNP regression coefficients that are unique to SZ, corrected for BIP. We refer to this set of summary statistics as "unique" $SZ_{(\text{min BIP})}$. We then repeated the look-up of the EA-associated lead SNPs in those summary statistics and note that the EA- and "unique" $SZ_{(\text{min BIP})}$ results have been derived from independent samples, similar to our main look-up described in Supplementary Note sections 1-3.

A GWIS infers genome-wide summary statistics for a (non-linear) function of phenotypes for which GWAS summary statistics are available[39]. Here, in particular, we wish to infer for each SNP the effect on SZ, conditioned upon its effect on BIP. One possible approximation involves a GWIS of the following linear regression function:

$$SZ = \beta * BIP + e$$

where the parameter β is estimated from the genetic covariance between SZ and BIP and the genetic variance in BIP as $\beta = \frac{cov_g(SZ,BIP)}{var_g(BIP)}$. The residual (e) is actually our trait of intrest, for which we use the term $SZ_{(\text{min BIP})}$. Using GWIS we infer the genome wide summary statistics for $SZ_{(\text{min BIP})}$ given the most recent PGC GWAS results for SZ (omitting the GRAS data collection)[2] and BIP[40]. The effect size with respect to $SZ_{(\text{min BIP})}$ for a single SNP is computed as:

$$eff_{SZ} - \beta * eff_{BIP} = eff_e$$

The standard error for each SNP effect is approximated using the delta method and accounts for the possible effect of sample overlap between the SZ and BIP GWAS.

As data input, we used the GWAS results on schizophrenia (excluding the GRAS data collection) described in Supplementary Note section 2. GWAS results for BIP[40] were obtained from the website of the PGC (https://www.med.unc.edu/pgc/files/resultfiles/pgc.cross.bip.zip).

9.2 Look-up in GWIS results schizophrenia – bipolar disorder

The "unique" $SZ_{(\text{min BIP})}$ results obtained from the GWIS were processed and merged with the EA GWAS results using the same procedures described in Supplementary Note section 3, leading to 1,153,214 SNPs that passed the quality control thresholds. This number is substantially lower than in the main look-up reported in Supplementary Note section 3 because the BIP GWAS was based on HapMap 2 imputation[41] not on 1000 Genomes imputation[42] like the SZ[2] and EA GWAS[1].

We repeated the clumping and look-up of the EA-associated lead SNPs in the cleaned and merged "unique" $SZ_{(\text{min BIP})}$ results following the steps described in Supplementary Note section 5.1.

346 approximately independent EA lead SNPs with $P_{EA} < 10^{-5}$ were identified. None of them was significantly associated with "unique" $SZ_{(\text{min BIP})}$ after Bonferroni correction ($P = \frac{0.05}{346} = 1.445 \times 10^{-4}$). Supplementary Figure 2 presents a Q-Q plot of this look-up. Supplementary Table 9.1 reports the full results.

9.2.1 Sign concordance

We compared the signs of the beta coefficients of the 346 EA lead SNPs ($P_{EA} < 10^{-5}$) with the beta coefficients of the "unique" $SZ_{(\text{min BIP})}$ results. If the signs were aligned, we assigned a "1" to the SNP and "0" otherwise. By chance, sign concordance is expected to be 50%. We tested if the observed sign concordance is different from 50% using the binomial probability test[14]. 154 of the 346 SNPs had the same sign (44.5%, $P = 0.046$, 2-sided). This result is consistent with a negative genetic correlation between the most strongly EA-associated SNPs and "unique" $SZ_{(\text{min BIP})}$ and contrasts with the positive genetic correlation between EA and SZ reported in Obkay et al[1].

9.2.1 Raw enrichment factor (not corrected for LD score of SNPs)

We calculated the raw enrichment factor of the EA-associated SNPs in the "unique" $SZ_{(\text{min BIP})}$ results using the approach described in Supplementary Note section 5.4.1. We obtained 109,188 approximately independent EA lead-SNPs.

For $P_{EA} < 10^{-5}$, we found 346 SNP ($\tau_{P_{EA}} = \frac{306}{109{,}188} = 0.317\%$). For $P_{SZ} < 0.05$, we found 6,190 SNP ($\tau_{P_{SZunique}} = \frac{6{,}190}{109{,}188} = 5.669\%$). Thus, under the null hypothesis of no enrichment we expect $\left[N_{S,EA \to SZunique}\right] = 109{,}188 \times 0.317\% \times 5.669\% = 20$ SNPs to have $P_{EA} < 10^{-5}$ and $P_{SZunique} < 0.05$ simulatenously. At these P value thresholds, we actually observe $N_{S,EA \to SZunique} = 32$ SNPs, implying a raw enrichment factor of $\frac{32}{20} = 1.6$. Thus, the observed enrichment of the EA-associated SNPs in the "unique" $SZ_{(\text{min BIP})}$ results is weaker than in our main look-up reported in Supplementary Note section 5.4.1, but it still deviates from the expectation under the null hypothesis.

9.2.2 Raw enrichment P value (not corrected for LD score of SNPs)

We repeated the procedures described in Supplementary Note section 5.4.2 and found a raw enrichment *P* value of 0.02 ($Z = 2.317$, 2-sided). Thus, although the enrichment of the EA-associated top SNPs is unlikely to be drawn from the same distribution as all "unique" $SZ_{(min\ BIP)}$ results with the same MAF distribution, the enrichment is weaker than in the main SZ GWAS results that did not control for the genetic overlap between SZ and BIP.

9.3 GWIS bipolar disorder - schizophrenia

Using the GWIS method and the data sources described above (Supplementary Note section 9.1), we also "purged" the genetic association results for BIP of their overlap with SZ, obtaining "unique" $BIP_{(min\ SZ)}$ results.

9.4 Genetic correlations of GWAS and GWIS results

To test if the genetic overlap of SZ with EA is partially due to their genetic correlation with other traits, we computed genetic correlations of SZ and EA with three other phenotypes of particular relevance—BIP, childhood IQ, and neuroticism. Given that SZ is sometimes referred to as a cognitive disorder[43,44], it is somewhat puzzling that previous studies did not find a significant (negative) genetic correlation between SZ and childhood IQ[33]. Furthermore, the personality trait neuroticism has been demonstrated to correlate across various psychiatric disorders and is positively associated ($r \approx 0.4$) with a general psychopathology factor (p)[45]. In addition, moderate and strong negative genetic correlations of neuroticism have been reported for EA[1] and depressive symptoms[4], respectively, raising the possibility that neuroticism may contribute to the genetic overlap between EA and SZ. Finally, Nieuwboer et al.[39] suggest that the genetic correlation between EA and SZ may be induced by the genetic correlation between SZ and BIP as well as the genetic correlation between BIP and EA. Extending this logic, it could also be that the lack of a clear negative genetic correlation between SZ and childhood IQ is induced by the genetic correlation between SZ and BIP.

Our analyses are facilitated by the fact that large-sample GWAS summary statistics are available in the public domain for all five traits (see data sources in Supplementary Table 9.2). In addition to analysing GWAS summary statistics, we also included the GWIS results described above in Supplementary Note sections 9.1 and 9.3 (GWIS $SZ_{(min\ BIP)}$ and GWIS $BIP_{(min\ SZ)}$).

Supplementary Table 9.2 and Figure 4 display the results. When using the GWAS summary statistics, we obtain results very closely resembling those in earlier studies. In particular, we find a strong positive genetic correlation between SZ and BIP ($r_g = 0.72$, $P = 8.57 \times 10^{-60}$)[33,39], a positive genetic correlations of EA with childhood IQ ($r_g = 0.74$, $P = 2.22 \times 10^{-30}$) and BIP ($r_g = 0.27$, $P = 1.75 \times 10^{-11}$), as well as a negative genetic correlation between EA and neuroticism ($r_g = -0.25$, $P = 7.08 \times 10^{-17}$)[1]. Also in line with previous reports is the insignificant genetic correlation between SZ and childhood IQ ($r_g = -0.03$, $P = 0.61$)[33] as well as the positive genetic correlation between SZ and EA ($r_g = 0.09$, $P = 1.04 \times 10^{-04}$)[1], both of which are counter-intuitive results. Furthermore, we find some positive genetic correlation of neuroticism with SZ ($r_g = 0.19$, $P = 4.5 \times 10^{-07}$) and BIP ($r_g = 0.10$, $P = 0.06$). However, these results are too weak to justify a GWIS approach that would purge the SZ and BIP results of their genetic correlation with neuroticism.

Interestingly, the genetic correlations change substantially when we purge the SZ results of their genetic overlap with BIP (GWIS $SZ_{(min\ BIP)}$) and vice versa (GWIS $BIP_{(min\ SZ)}$). The

genetic correlations between EA and IQ with $SZ_{(\text{min BIP})}$ are now negative and significant (r_g = -0.16, $P = 3.88\times10^{-04}$ and r_g = -0.31, $P = 6.00 \times 10^{-03}$ respectively), which is more in line with the idea of SZ being a cognitive disorder.[43] Furthermore, the genetic correlations of EA and IQ with $BIP_{(\text{min SZ})}$ remain positive and get somewhat stronger (r_g = 0.31, $P = 2.87 \times 10^{-07}$ and r_g = 0.33, $P = 3.18 \times 10^{-02}$ respectively). The genetic correlations of $SZ_{(\text{min BIP})}$ and $BIP_{(\text{min SZ})}$ with neuroticism, however, remain quite stable.

These results suggest two things: First, the genetic overlap between SZ and EA reported in Supplementary Note section 5 and the small, positive genetic correlation between the two traits is indeed to some extent caused by their genetic overlap with both BIP and childhood IQ (and not by their overlap with neuroticism). Second, once we purge the genetic association with SZ of their overlap with BIP, we see that the remaining part of "unique" $SZ_{(\text{min BIP})}$ has *negative* genetic correlations with childhood IQ and EA. Thus, SZ diagnoses that have been used in large-scale GWAS analyses until now seem to comprise of at least two subtypes of the disease that have different genetic components: One part resembles a cognitive disorder which does not overlap with BIP, and one part does overlap with BIP but is not characterised by cognitive deficiencies.

10 Simulating assortative mating

Previous studies found substantial assortative mating for EA, with spousal correlations in the range between 0.45 and 0.66, which led to a genetic resemblance among spouses for EA-associated genetic markers[46]. There is also evidence for substantial assortative mating for psychiatric disorders (in particular, SZ) with spousal correlations around 0.4[47]. One possibility is that strong, simultaneous assortative mating on EA and SZ may cause an enrichment of EA-associated loci for SZ. If this happens in the absence of phenotypic and genetic correlations between the two traits, one may call such an enrichment "spurious" because the genetic variants for EA would have no actual influence on SZ, even if they would be found to be robustly associated with SZ. We ran simulations to test how likely it is that our results may be driven by strong, assortative mating.

10.1 Simulations

10.1.1 Description

We simulated an initial generation of n = 25,000. For each individual, two sets of 5,000 genetic markers were drawn from a binomial distribution with a 50% chance to be either 0 or 1. The sum of the two copies of each marker is the genotype of the individual. We assumed that two non-overlapping subsets of markers (500 each) were causal for either EA or SZ. We further assumed that both EA and SZ are 100% heritable, binary outcomes. The frequency of high educational attainment was set to 0.3, and the frequency of SZ to 0.2. The phenotype of each individual was determined by examining the value of a polygenic score based on the known causal markers for each trait. We determined the relevant cut-off point of the score such that it matched the assumed frequencies of both traits in the population.

The initial generation went through a matching algorithm where each person was matched to a spouse and spousal correlations for both traits was assumed to be 0.6. Next, each couple had exactly two offspring and each offspring's genotype was drawn from the genotypes of the parents. The offspring generation was then also matched to a spouse and the process was

repeated for 50 generations. After each generation, we tested the causal EA markers for association with SZ and computed the raw enrichment test described in Supplementary Note section 5.2.

Our assumptions were not necessarily meant to represent the real world. Instead, our aim was to simulate a set-up with relatively extreme assumptions about heritability, assortative mating, and the prevalence of SZ, all of which would increase our chance of finding spurious enrichment. However, we note that our model was also deliberately simple: Implicitly, we assumed that there is no selective pressure on any genetic marker and that the effect sizes of the markers are constant over time. In reality, this may not be the case and both the phenotype and its genetic architecture may evolve.

10.1.2 Power

The power of our enrichment test is a function of the power to detect associations of individual genetic markers with SZ. To calculate power, we assumed an *odds* ratio of 1.117. This *odds* ratio was calculated using the fact that all causal markers had equal effect sizes and individual markers were drawn from a Bernoulli distribution, such that the polygenic score can be seen as a draw from a binomial distribution with parameters $n = 1{,}000$ and $P = 0.5$. As described above, this score was used to determine the SZ type. Specifically, a simulated individual had SZ if the score was larger than 513, which gives an overall chance of 20 percent to get SZ. The *odds* ratio was calculated using the following formula:

$$P_1 = P(\,SZ = 1 \mid X_{SZ} = 1), \qquad P_0 = P(\,SZ = 1 \mid X_{SZ} = 0)$$

$$Odds = \frac{P_1(1 - P_0)}{P_0(1 - P_1)}$$

Where P_1 is the chance of getting SZ given that you have at least one of the causal variants increasing the chance of SZ, X_{SZ}, and P_0 is the chance of getting SZ given that you have the opposite variant. P_1 and P_0 can calculated from the binomial distribution. P_1 is the chance that a draw from a binomial distribution, with parameters $n = 999$ and $P = 0.5$, is larger than 512. P_0 is the chance that a draw from the same distribution is larger than 513.

We had 93.3% power to detect the causal markers associated with SZ at nominal significance and 18.1% after Bonferroni correction ($P = \frac{0.05}{5000} = 10^{-5}$). Furthermore, we had 80% power to detect (spurious) effects of $Odds \geq 1.093$ at $P = 0.05$ and of $Odds \geq 1.182$ after Bonferroni correction. The raw enrichment test had even more power because it took the entire distribution of tested *P* values into account, not only the effects of one single SNP. Thus, we were well-powered to detect spurious enrichment in our simulation, even if the effects of individual genetic markers are relatively small.

10.2 Results

If assortative mating would cause the genetic overlap between EA and SZ, one would expect the absolute value of the *Z*-statistic from the enrichment test to increase over time. At some point, the difference between the two subsets would become statistically significant. The *Z*-statistic of the test for each generation is plotted in Supplementary Figure 3. There is no persistent trend in the *Z*-statistics over time and the mean of the *Z*-statistics is not significantly different from 0 ($P = 0.776$). Furthermore, we cannot reject the hypothesis that the *Z*-statistics are simply drawn from a standard normal distribution (Shapiro-Wilk test, $P = 0.075$, one-sided). Nevertheless, the *Z*-statistic drops below -1.96 in two simulated

generations, indicating that the causal EA markers have lower *P* values than the non-causal markers. Yet, this result does not persist over time. In conclusion, it may be possible to find (spurious) enrichment in some generations by chance, but it is unlikely that assortative mating is a major cause for the strong genetic overlap that we observed between EA and SZ.

11 Biological annotation

11.1 Prioritisation of genes, pathways, and tissues/cell types with DEPICT

To gain first insights into possible biological pathways that are indicated by our genetic associations, we applied Data-driven Expression Prioritized Integration for Complex Traits (DEPICT). DEPICT is a novel data-driven integrative method that uses reconstituted gene sets based on massive numbers of experiments measuring gene expression to (1) prioritise genes and gene sets and (2) identify tissues and cell types were prioritised genes are highly expressed. The method has been described in detail elsewhere[1,4]. The input for our analyses (DEPICT version 1 release 194) were the 132 EA lead-SNPs that are also nominally associated with SZ.

Significant reconstituted gene sets

DEPICT identified 111 significant reconstituted gene sets at an FDR below 5% (Supplementary Table 10.1). To identify independent biological groupings, we computed the pairwise Pearson correlations of all significant gene sets using the "network_plot.py" script provided with DEPICT. Next, we used the Affinity Propagation method on the Pearson distance matrix to cluster the findings[48]. The Affinity Propagation method automatically chooses an exemplar for each cluster (Supplementary Table 10.2). Figure 5.a visualises the results of this analyses, showing only one exemplar per gene sets (n = 19). We briefly describe the implicated gene sets below. The definitions are taken from AmiGO[49], the Mouse Genome Database[50], the Reactome pathway Knowledgebase[51] and GeneCards[52].

npBAF complex (7-set cluster) is named after the GO cellular component (GO:0071564) defined as "A SWI/SNF-type complex that is found in neural stem or progenitor cells". The prefix *np* stands for *neural progenitor*. The npBAF complex is essential for the self-renewal/proliferative capacity of multipotent neural stem cells.

Transcription cofactor activity (12-set cluster) is named after the GO molecular function (GO:0003712) defined as "Interacting selectively and non-covalently with a regulatory transcription factor and also with the basal transcription machinery in order to modulate transcription. Cofactors generally do not bind the template nucleic acid, but rather mediate protein-protein interactions between regulatory transcription factors and the basal transcription machinery".

REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES (15-set cluster) is named after the Reactome pathway centred on genes involved in transmission across chemical synapses. Chemical synapses are specialised junctions that are used for communication between neurones, neurones and muscle or gland cells. The pre-synaptic neurone communicates via the release of neurotransmitter which binds the receptors on the post-synaptic cell.

Dendrite (21-set cluster) is named after a large and heterogeneous GO cellular component (GO: 0030425) defined as "A neuron projection that has a short, tapering, often branched,

morphology, receives and integrates signals from other neurons or from sensory stimuli, and conducts a nerve impulse towards the axon or the cell body. In most neurones, the impulse is conveyed from dendrites to axon via the cell body, but in some types of unipolar neurone, the impulse does not travel via the cell body".

Abnormal cerebral cortex morphology (5-set cluster) is named after the Mammalian Phenotype category (MP:0000788) defined as "any structural anomaly of a thin layer of grey matter on the surface of the cerebral hemisphere that folds into gyri".

Dendritic spine morphogenesis (3-set cluster) is named after the GO biological process (GO:0060997) defined as "The process in which the anatomical structures of a dendritic spine are generated and organised. A dendritic spine is a protrusion from a dendrite and a specialised subcellular compartment involved in synaptic transmission".

REACTOME_AXON_GUIDANCE (4-set cluster) is named after the Reactome pathway centred on genes involved in axon guidance. Axon guidance or axon pathfinding is the process by which neurones send out axons to reach the correct targets.

GRIN2A PPI subnetwork (9-set cluster) is named after the gene *GRIN2A* (glutamate ionotropic receptor NMDA-type subunit 2A). This gene encodes a member of the glutamate-gated ion channel protein family. The encoded protein is an N-methyl-D-aspartate (NMDA) receptor subunit. The most significantly enriched member set is "protein serine/threonine phosphatase complex", which is named after GO cellular component (GO:0008287) defined as "A complex, normally consisting of a catalytic and a regulatory subunit, which catalyzes the removal of a phosphate group from a serine or threonine residue of a protein".

SNW1 PPI subnetwork (3-set cluster) is named after the gene *SNW1* (SNW Domain Containing 1) encodes a coactivator that enhances transcription from some Pol II promoters.

DLGAP3 PPI subnetwork (3-set cluster) is named after the gene *DLGAP3* (Discs Large Homolog Associated Protein 3) that plays a role in the molecular organisation of synapses and neuronal cell signalling.

Neurone recognition (3-set cluster) is named after the GO biological process (GO:0008038) defined as "The process in which a neuronal cell in a multicellular organism interprets its surroundings".

WRB PPI subnetwork (3-set cluster) is named after the gene *WRB* (Tryptophan Rich Basic Protein) also known as *CHD5* (Congenital Heart Disease 5) that has a potential role in the pathogenesis of Down syndrome congenital heart disease.

Site of polarised growth (3-set cluster) is named after the GO cellular component (GO:0030427) defined as "Any part of a cell where non-isotropic growth takes place".

Central nervous system neurone axonogenesis (3-set cluster) is named after the GO biological process (GO:0021955) defined as "Generation of a long process from a neurone whose cell body resides in the central nervous system. The process carries efferent (outgoing) action potentials from the cell body towards target cells".

Regulation of neurone projection development (6-set cluster) is named after the GO biological process (GO:0010975) defined as "Any process that modulates the rate, frequency or extent of neurone projection development. Neurone projection development is the process whose specific outcome is the progression of a neurone projection over time, from its formation to the mature structure. A neurone projection is any process extending from a neural cell, such as axons or dendrites (collectively called neurites)".

WDR37 PPI subnetwork (7-set cluster) is named after the gene *WDR37* (WD Repeat Domain 37) that encodes a member of the WD repeat protein family and may facilitate the formation of heterotrimeric or multiprotein complexes.

REACTOME_PI3K_EVENTS_IN_ERBB4_SIGNALING (2-set cluster) is named after the Reactome pathway centred on genes involved in PI3K events in ERBB4 signalling. ERBB4 is a member of the Tyr protein kinase family and the epidermal growth factor receptor subfamily. It is for the normal development of the embryonic central nervous system, especially for normal neural crest cell migration and normal axon guidance.

Regulation of skeletal muscle fibre development (1-set cluster) is named after the GO biological process (GO:0048742) defined as "Any process that modulates the frequency, rate or extent of skeletal muscle fibre development".

HMGB2 PPI subnetwork (1-set cluster) contains only one single gene set. It is named after the gene *HMGB2* (High Mobility Group Box 2) that encodes a member of the non-histone chromosomal high mobility group protein family.

Significant tissue/cell types

DEPICT determines the enrichment of expression in particular tissues and cell types by testing whether the genes overlapping the GWAS loci are highly expressed in any of 209 Medical Subject Heading (MeSH) annotations. Interestingly, all the significantly enriched tissues (FDR < 0.05) are related to the nervous system except retina, which is annotated to sense organs (Fig. 5.b). Furthermore, we observed only 1 significantly enriched cell-type, namely "*Neural Stem Cells*". All the significantly enriched tissues and cell-type along with the gene names are listed in Supplementary Table 10.3.

Significant gene prioritisation

Any particular locus centred on a SNP may contain multiple genes. One straightforward approach is to nominate a gene that is closest to the SNP. But this approach does not consider if the expression of the gene is likely to be altered or regulated by the causal site in the locus. Therefore, we used DEPICT to map genes to associated loci, which prioritise important genes that share similar annotations in bioinformatic databases. For our 132 lead SNPs, DEPICT significantly prioritized (FDR < 0.05) 56 genes (Supplementary Table 10.4). For the two novel SNPs reported in this study (rs7336518 and rs7522116), DEPICT points to the *FOXO6* (Forkhead Box O6) and the *SLITRK1* (SLIT and NTRK Like Family Member 1) genes. *FOXO6* is predominantly expressed in the hippocampus and has been suggested to be involved in memory consolidation, emotion and synaptic function[53,54]. Similarly, *SLITRK1* is also highly expressed in the brain[55], particularly in the frontal lobe, and has previously been suggested as a candidate gene for neuropsychiatric disorders[56]. In particular, *SLITRK1* is also associated with Tourette syndrome, which is characterised by persistent involuntary vocal and motor tics and often occurs together with Obsessive-Compulsive disorder and ADHD[57–59].

11.2 GWAS catalog lookup

In order to investigate the novelty of the 21 SNP associations that were found significant for SZ after Bonferroni correction, reported in **Table 1**, we performed a lookup in the GWAS catalogue with the SNPs and all their "LD partners" (i.e. all SNPs with an $r^2 > 0.5$ within a 250kb window). The LD partners were extracted with PLINK[5] using a version of the 1000G

reference panel specifically harmonised to combine 1000G phase 1 and phase 3 imputed data[60], and the reference panel has been described previously[4]. The result of the GWAS catalogue lookup is reported in Supplementary Table 10.5.

We searched the GWAS catalog[61] (revision 2016-08-25, downloaded on 2016-08-29)[c] to see if any of the associated SNPs or their LD partners have been reported to be associated with a phenotype previously. 13 out of the 21 SNPs (or their LD partners) have reported associations with SZ in the GWAS catalogue – eight out of the 21 SNPs have no reported associations with SZ. Combining the associations reported in the GWAS catalog and those reported in Ripke et al.[2], Pardinal et al.[8], and Hellard et al.[9], we find two SNPs that have not previously been found to be associated with SZ at genome-wide significance ($P < 5\times10^{-8}$) in any of these sources – rs7522116 and rs7336518.

[c] URL: https://www.ebi.ac.uk/gwas/api/search/downloads/full

12 References

1 Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**: 539–542.

2 Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**: 421–427.

3 Rietveld CA, Esko TT, Davies G, Pers TH, Turley PA, Benyamin B *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci U S A* 2014; **111**: 13790–13794.

4 Okbay A, Baselmans BML, Neve J-E De, Turley P, Nivard MG, Fontana MA *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* doi:10.1038/ng.3552.

5 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

6 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.

7 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.

8 Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. 2016http://dx.doi.org/10.1101/068593.

9 Le Hellard S, Wang Y, Witoelar A, Zuber V, Bettella F, Hugdahl K *et al.* Identification of gene loci that overlap between schizophrenia and educational attainment. *Schizophr Bull* 2016. doi:10.1093/schbul/sbw085.

10 Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet Part B Neuropsychiatr Genet* 2015; **168**: 649–659.

11 Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science (80-)* 2013; **340**: 1467–1471.

12 Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. Converting among effect sizes. In: *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd.: Chichester, United Kingdom, 2009, pp 45–51.

13 Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; **44**: 247–250.

14 Sheskin D. The binomial sign test for a single sample. In: *Handbook of Parametric and Nonparametric Statistical Procedures*. Taylor & Francis Group: Boca Raton, 2007, pp 289–311.

15 Nachar N. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Quant Methods Psychol* 2008; **4**: 13–20.

16 Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–295.

17 Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for

human height and body mass index. *Nat Genet* 2015; **47**: 1114–20.

18 Gazal S, Finucane H, Furlotte NA, Loh P-R, Palamara PF, Liu X *et al.* Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection. 2016 doi:http://dx.doi.org/10.1101/082024.

19 Begemann M, Grube S, Papiol S, Malzahn D, Krampe H, Ribbe K *et al.* Modification of cognitive performance in schizophrenia by complexin 2 gene polymorphisms. *JAMA Psychiatry* 2010; **67**: 879–888.

20 Ribbe K, Friedrichs H, Begemann M, Grube S, Papiol S, Kästner A *et al.* The cross-sectional GRAS sample: A comprehensive phenotypical data collection of schizophrenic patients. *BMC Psychiatry* 2010; **10**: 91.

21 Hammer C, Zerche M, Schneider A, Begemann M, Nave K-A, Ehrenreich H. Apolipoprotein E4 carrier status plus circulating anti-NMDAR1 autoantibodies: Association with schizoaffective disorder. *Mol Psychiatry* 2014; **19**: 1054–6.

22 Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10**: 5–6.

23 Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.

24 Lehrl S. *Mehrfachwahl-Wortschatz-Intelligenztest: MWT-B.* Spitta: Balingen, 1999.

25 Kay SR, Flszbein A, Opfer LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. .

26 Association AP. *Diagnostic and statistical manual of mental disorders DSM-IV*. American Psychiatric Publishing, Inc., 1994.

27 Guy W. Clinical Global Impression (CGI). In: *ECDEU Assessment manual for psychopharmacology*. National Institute of Health, Psycho-pharmacology Research Branch.: Rockville, MD., 1976, pp 218–222.

28 Statistics UI for. International Standard Classification of Education. 2006.http://www.uis.unesco.org/Library/Documents/isced97-en.pdf (accessed 1 Jan2015).

29 Davis JM, JM D, Appleton WS DJ, TH B, LE H, GL K *et al.* Comparative doses and costs of antipsychotic medication. *Arch Gen Psychiatry* 1976; **33**: 858.

30 Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 2011; **43**: 977–83.

31 Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Massachusetts Institute of Technology: Cambridge, MA, 2002 doi:10.1515/humr.2003.021.

32 Stata C. Stata Statistical Software: Release 14. 2015.

33 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Consortium R *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**: 1236–1241.

34 Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.

35 Levine SZ, Rabinowitz J. A population-based examination of the role of years of education, age of onset, and sex on the course of schizophrenia. *Psychiatry Res* 2009; **168**: 11–17.

36 Bambole V, Johnston M, Shah N, Sonavane S, Desouza A, Shrivastava A. Symptom overlap between schizophrenia and bipolar mood disorder: Diagnostic issues. *Open J Psychiatry* 2013; **3**: 8–15.

37 Daneluzzo E, Arduini L, Rinaldi O, Di Domenico M, Petruzzi C, Kalyvoka A *et al.* PANSS factors and scores in schizophrenic and bipolar disorders during an index acute episode: a further analysis of the cognitive component. *Schizophr Res* 2002; **56**: 129–136.

38 Wooldridge JM. Multiple Regression Analysis: Estimation. In: *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2013, pp 70–76.

39 Nieuwboer HA, Pool R, Dolan CV, Boomsma DI, Nivard MG. GWIS: Genome-wide inferred statistics for functions of multiple phenotypes. *Am J Hum Genet* 2016; **99**: 917–927.

40 Consortium C-DG of the PG. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013; **381**: 1371–1379.

41 Consortium TIH. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.

42 The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.

43 Kahn RS, Keefe RSE, JD H, B E, GM K, H D *et al.* Schizophrenia is a cognitive illness. *JAMA Psychiatry* 2013; **70**: 1107.

44 Kraepelin E. *Psychiatrie: Ein Lehrbuch für Studierende und Ärzte*. 4th ed. Verlag von Johann Ambrosius Barth: Leipzig, Germany, 1893.

45 Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S *et al.* The p Factor. *Clin Psychol Sci* 2014; **2**: 119–137.

46 Hugh-Jones D, Verweij KJH, St. Pourcain B, Abdellaoui A. Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence* 2016; **59**: 103–108.

47 Nordsletten AE, Larsson H, Crowley JJ, Almqvist C, Lichtenstein P, Mataix-Cols D *et al.* Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* 2016; **73**: 354.

48 Frey BJ, Dueck D. Clustering by passing messages between data points. *Science (80-)* 2007; **315**.

49 Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* 2009; **25**: 288–9.

50 Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* 2015; **43**: D726–D736.

51 Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* 2016; **44**: D481–D487.

52 Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S *et al.* The GeneCards Suite: From gene data mining to disease genome sequence analyses. *Curr Protoc Bioinforma* 2016; : 1.30.1-1.30.33.

53 Salih DAM, Rashid AJ, Colas D, de la Torre-Ubieta L, Zhu RP, Morgan AA *et al.* FoxO6 regulates memory consolidation and synaptic function. *Genes Dev* 2012; **26**: 2780–2801.

54 Maiese K. FoxO Proteins in the Nervous System. *Anal Cell Pathol* 2015; **2015**: 1–15.

55 Aruga J, Yokota N, Mikoshiba K. Human SLITRK family genes: genomic organization and expression profiling in normal brain and brain tumor tissue. *Gene* 2003; **315**: 87–94.

56 Proenca CC, Gao KP, Shmelkov S V, Rafii S, Lee FS, Aruga J *et al.* Slitrks as emerging candidate genes involved in neuropsychiatric disorders. *Trends Neurosci* 2011; **34**: 143–53.

57 Carter AS, O'Donnell DA, Schultz RT, Scahill L, Leckman JF, Pauls DL. Social and emotional adjustment in children affected with Gilles de la Tourette's syndrome: associations with ADHD and family functioning. Attention Deficit Hyperactivity Disorder. *J Child Psychol Psychiatry* 2000; **41**: 215–23.

58 Lombroso PJ, Scahill L. Tourette syndrome and obsessive–compulsive disorder. *Brain Dev* 2008; **30**: 231–237.

59 O'Rourke JA, Scharf JM, Yu D, Pauls DL. The genetics of Tourette syndrome: A review. *J Psychosom Res* 2009; **67**: 533–545.

60 Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

61 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001-1006.

62 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA *et al.* Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356–369.

13 Additional acknowledgments

All individuals who participated in the study provided informed consent (see section 6 and references[1,2]).

GRAS data collection: The research of EU-AIMS receives support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115300, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013), from the EFPIA companies and from Autism Speaks. We thank all subjects for participating in the study, and all the many colleagues who have contributed over the past decade to the GRAS data collection.

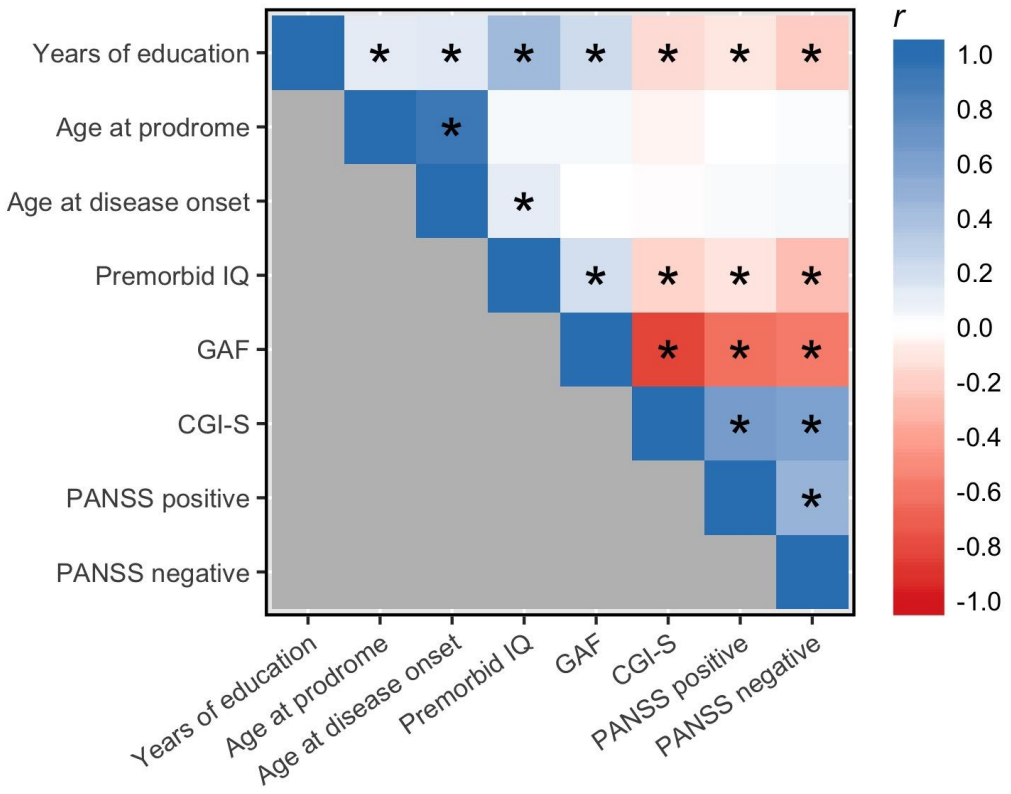
Niels Rietveld gratefully acknowledges funding from the Netherlands Organization for Scientific Research (NWO Veni grant 016.165.004).

13.1 Individual author contributions:

			Conceptualization	Data Curation	Formal Analysis	Methodology	Project Administration	Resources	Software	Supervision	Visualization	Writing - Original Draft Preparation	Writing - Review & Editing
Vikas		Bansal			X	X					X	X	X
Marina		Mitjans			X	X					X	X	X
Casper	AP	Burik			X	X					X	X	
Martin		Begemann		X									
Stefan		Bonn						X		X			
Richard		Karlsson Linnér			X						X		
Aysu		Okbay		X	X								
Cornelius	A	Rietveld									X		
Stephan		Ripke		X	X								
Michel		Nivard			X	X			X	X	X		X
Hannelore		Ehrenreich	X	X		X	X	X		X		X	X
Philipp	D	Koellinger	X	X	X	X	X	X		X		X	X

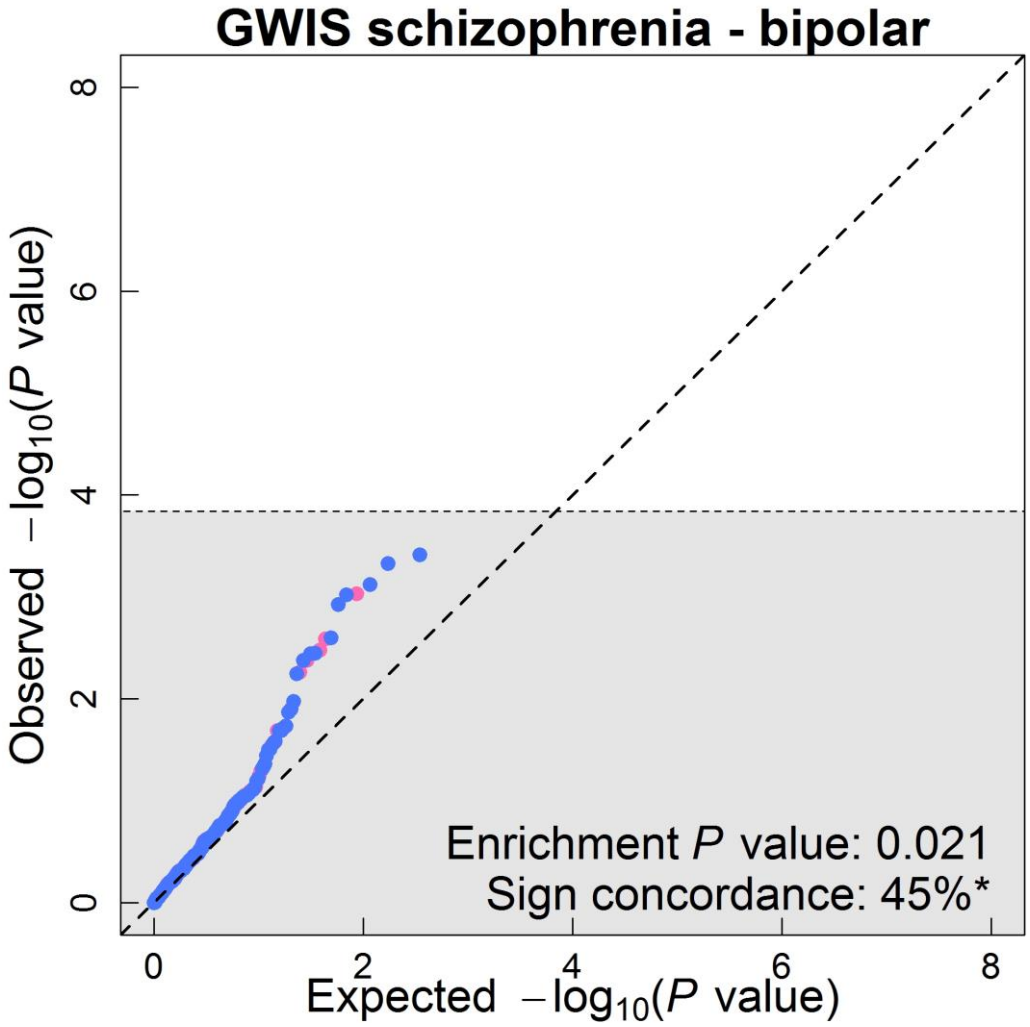
14 Supplementary Figures

Supplementary Figure 1: Phenotypic correlations in the GRAS data collection, SZ cases only



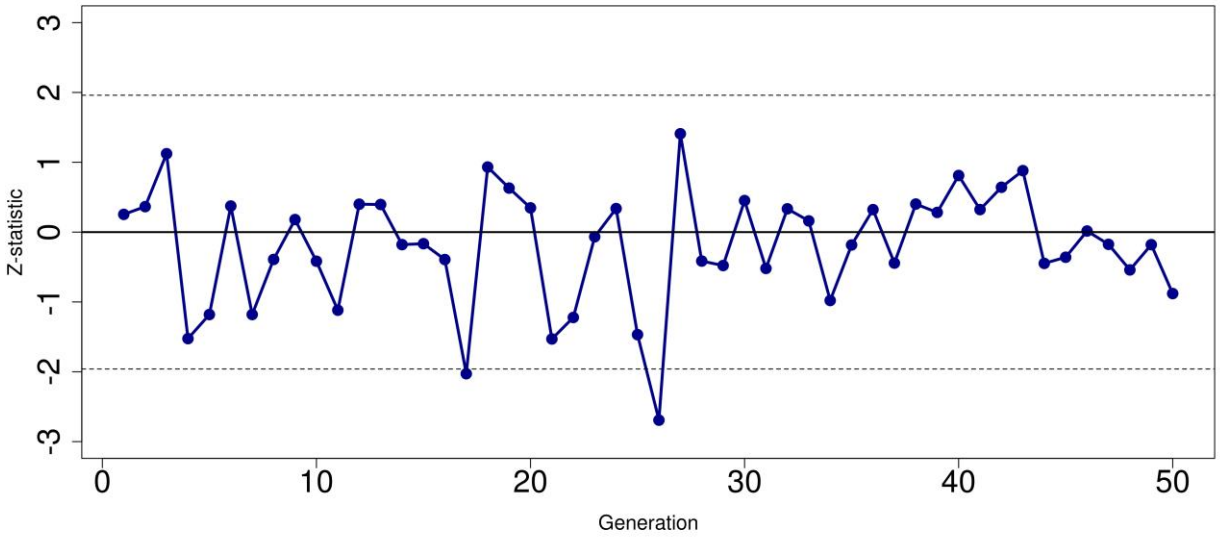
Notes: Phenotypic correlations (Pearson's *r*) among schizophrenia patients in the GRAS data collection. The direction of the correlations is indicated by the coloring (blue for positive, red for negative) and the magnitude of the correlations is indicated by the gradient of the color. Black dots indicate nominal $P < 0.01$. GAF: Global Assessment of Functioning. CGI-S: Clinical Global Impression of Severity. PANSS: Positive and Negative Syndrome Scale. See detailed statistics in Supplementary Table 8.1. and detailed description of the measures in Supplementary Section 6.

Supplementary Figure 2: **Q–Q plot of the 346 EA-associated SNPs for "unique" $SZ_{(min\ BIP)}$.**



Notes: SNPs with concordant effects on both phenotypes are pink, and SNPs with discordant effects are blue. SNPs outside the grey area would have passed the Bonferroni-corrected significance threshold that corrects for the total number of SNPs we tested ($P < 0.05/346 = 1.445\times10^{-4}$). Observed and expected *P* values are on the $-\log_{10}$ scale. For the sign concordance test: $P = 0.046$, 2-sided.

Supplementary Figure 3: Results of the assortative mating simulations – raw enrichment *Z*-statistics of causal SNPs for EA on SZ



Notes: The graph shows the *Z*-statistic of the Wilcoxon rank sum test (i.e. the raw enrichment *P* value according to the methods described in Supplementary Note section 5) for each simulated generation. Causal genetic markers for EA and non-causal markers are tested for association with SZ. *Z*-statistic of 1.96 is shown by dashed lines.