Novel loci associated with Attention-deficit/hyperactivity disorder are revealed by leveraging polygenic overlap with Educational Attainment
RH: Novel loci associated with ADHD

Alexey A. Shadrin, PhD; Olav B. Smeland, MD, PhD; Tetyana Zayats, MD, PhD; Andrew J. Schork, PhD; Oleksandr Frei, PhD; Francesco Bettella, PhD; Aree Witoelar, PhD; Wen Li, PhD; Jon A. Eriksen, PhD; Florian Krull, PhD; Srdjan Djurovic, MD, PhD; Stephen V. Faraone, MD, PhD; Ted Reichborn-Kjennerud, MD, PhD; Wesley K. Thompson, PhD; Stefan Johansson, MD, PhD; Jan Haavik, MD, PhD; Anders M. Dale, PhD; Yunpeng Wang, PhD; Ole A. Andreassen, MD, PhD

Drs. Shadrin, Smeland, Frei, Bettella, Witoelar, Li, Eriksen, Krull, Wang and Andreassen are with NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway and Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. Dr. Wang is also with Department of Neurosciences, University of California, San Diego, La Jolla, CA. Drs. Zayats, Johansson and Haavik are with K.G. Jebsen Centre for Neuropsychiatric Disorders, Department of Clinical Science, University of Bergen, Bergen, Norway. Dr. Johansson is also with Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway. Dr. Haavik is also with Division of Psychiatry, Haukeland University Hospital, Bergen, Norway. Dr. Schork is with Department of Cognitive Science, University of California, San Diego, La Jolla, CA and Institute of Biological Psychiatry, Medical Health Center, Sct. Hans Hospital and University of Copenhagen, Copenhagen, Denmark. Dr. Djurovic is with Department of Medical Genetics, Oslo University Hospital, Oslo, Norway and NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen, Norway. Dr Faraone is with KG Jebsen Centre for Neuropsychiatric Disorders, Department of Biomedicine, University of Bergen, Bergen, Norway, Departments of Psychiatry and Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, New York. Dr Reichborn-Kjennerud is with Division of Mental Health, Norwegian Institute of Public Health, Oslo, Norway and Institute of Clinical Medicine, University of Oslo, Oslo, Norway. Dr. Thompson is with Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA. Dr Dale is with NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, Department of Neurosciences, University of California, San Diego, La Jolla, CA, Department of Radiology, University of California, San Diego, La Jolla, CA, Department of Psychiatry, University of California, San Diego, La Jolla, CA.

Correspondence to Alexey A. Shadrin, PhD, Oslo Universitetssykehus HF, Ullevål, Avdeling for psykoseforskning, Kirkeveien 166, 0424, Oslo, Norway; e-mail: a.a.shadrin@medisin.uio.no

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Abstract

Objective: Attention-deficit/hyperactivity disorder (ADHD) is a common and highly heritable psychiatric condition. By exploiting the reported relationship between ADHD and educational attainment (EA), we here aimed to improve discovery of ADHD-associated genetic variants and investigate genetic overlap between these phenotypes.

Method: A conditional/conjunctional false discovery rate (condFDR/conjFDR) method was applied to genome-wide association study (GWAS) data on ADHD (2064 trios, 896 cases and 2455 controls) and EA ($N = 328917$) to identify ADHD-associated loci and loci overlapping between ADHD and EA. Identified single nucleotide polymorphisms (SNPs) were tested for association in an independent population-based study of ADHD symptoms ($N = 17666$). Genetic correlation between ADHD and EA was estimated using LD score regression and Pearson correlation.

Results: At levels of condFDR < 0.01 and conjFDR < 0.05 we identified five ADHD-associated loci, three of these being shared between ADHD and EA. None of these loci had been identified in the primary ADHD GWAS, demonstrating the increased power provided by the condFDR/conjFDR analysis. Leading SNPs for 4 of 5 identified regions are in introns of protein coding genes: *KDM4A, MEF2C, PINK1, RUNX1T1,* while the remaining one is an intergenic SNP on chromosome 2 at 2p24. Consistent direction of effects in the independent study of ADHD symptoms was shown for 4 of 5 identified loci. A polygenic overlap between ADHD and EA was supported by significant genetic correlation ($r_g = -0.403$, $p = 7.90 \times 10^{-8}$) and >10-fold mutual enrichment of SNPs associated with both traits.

Conclusion: We identified five novel loci associated with ADHD and provided evidence for a shared genetic basis between ADHD and EA. These findings could aid understanding the genetic risk architecture of ADHD and its relation to EA.

2

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a common neurodevelopmental condition, caused by interplay of genetic and environmental risk factors. Its prevalence is estimated to be 5% in school-aged children and 2.50% in adults[1]. The heritability of ADHD is one of the highest reported among psychiatric disorders in epidemiological studies, estimated at 0.70–0.80[1]. However, it has been difficult to identify genetic risk variants that account for the high heritability of ADHD, resulting in a relatively modest SNP-based heritability, currently estimated at 0.28[2]. This may be in part explained by its complex phenotypic structure (heterogeneous clinical features, developmental course and outcome, high rate of comorbid symptoms and disorders[3]) and genetic architecture with a highly polygenic etiology, with both common and rare variants contributing small effects[4]. Moreover, large sample sizes are needed for reliable detection of such effects. The relatively small samples of existing ADHD genetic studies, as compared to those available for other psychiatric disorders[5,6], present an additional challenge. Up to now, no published GWASs have been able to detect genome-wide significant association ($p < 5.00 \times 10^{-8}$) for ADHD.

It is well-established that complex traits often have a polygenic structure with shared genetic background[7,8]. Recently, a conditional/conjunctional false discovery rate (condFDR/conjFDR) method was developed[9] to exploit overlapping association across GWASs and thereby boost association signals in GWAS of one phenotype by combining it with genome-wide association data of another phenotype (condFDR) or enable detection of specific genetic loci shared between two phenotypes (conjFDR). If genetic overlap between two phenotypes exists, the method offers for increased statistical power compared to conventional multiple hypotheses testing approaches[10,11]. This

3

method was successfully applied to discover novel associations and to detect shared genetic variants in various complex disorders, including neurological[12,13] and psychiatric[9] diseases.

ADHD is consistently associated with lower levels of EA[1,14]: the percentage of US adolescents not completing high school is 5%, whereas it is approximately 35% for adolescents diagnosed with ADHD[15]. There are several ways in which ADHD may relate to lower EA, which are not mutually exclusive. First, the clinical and cognitive symptoms of ADHD (e.g. attention deficits) may directly perturb EA. Secondly, ADHD has a number of common comorbidities, including learning disabilities[16], mood disorders[16] and disruptive behavior[16], associated with lower EA. Another possibility is that ADHD and EA share causative factors. Recent findings demonstrate negative genetic correlation between ADHD and EA ($r_g = -0.305$, se $= 0.141$, p $= 3.00 \times 10^{-2}$)[17], suggesting that genetic variants conferring risk to ADHD may contribute to lower EA in the general population. Thus, we can hypothesize that ADHD and EA may have a shared genetic basis and may amplify association signal by combining these phenotypes in condFDR/conjFDR method.

In contrast to ADHD, where the currently published largest GWASs contain less than 4000 cases[18,19], the latest GWAS on EA contains more than 300000 individuals, uncovering multiple genome-wide significantly associated loci[20]. Combining this EA GWAS with moderately-powered GWAS of ADHD[18] in the condFDR/conjFDR approach, we aimed here at identifying novel loci associated with ADHD as well as loci shared between ADHD and EA. The latter may provide insights into the molecular genetic mechanisms jointly influencing ADHD and EA and inform their biological underpinnings. Applying novel statistical methods, we also tested whether the observed phenotypic correlation between ADHD and EA implies a genetic correlation between these traits. Additionally, for the identified ADHD-associated variants, we assessed consistency of effect directions in an independent population based study of ADHD symptoms and performed *in silico* analyses of their functional effects (eQTL, expression quantitative loci).

4

Material and methods

Participant samples

We used ADHD data from the Psychiatric Genomics Consortium (PGC)[18]. The data set contains information from 2064 trios, 896 cases and 2455 controls. EA data were obtained from the Social Science Genetic Association Consortium (SSGAC)[20], where EA was measured as the number of years of schooling completed that was harmonized between different educational systems. For our analyses, we used summary statistics generated by the meta-analysis of all discovery and replication cohorts, except the 23andMe sample (64 datasets with total N = 328917).

Top association signals identified in our analyses were examined in the summary statistics from an independent GWAS of ADHD symptoms performed by EArly Genetics and Lifecourse Epidemiology (EAGLE) consortium[21]. Unlike the PGC case-control ADHD GWAS, EAGLE GWAS represents a meta-analysis of 9 population-based pediatric cohorts containing information on 17666 children under the age of 13 years with measures of ADHD symptom scores.

Detailed description of data used for analysis and data preprocessing steps is given in the supplemental material available online.

Statistical analyses

To assess genetic overlap between ADHD and EA and thus warrant subsequent condFDR/conjFDR analysis, we generated conditional QQ plots and fold-enrichment plots in both directions: conditioning ADHD on EA and vice versa[9]. To explore the nature of the polygenic overlap and test the hypothesis that the investigated phenotypes correlate genetically, we calculated Pearson correlations between association z-scores of ADHD and EA SNPs within nested subset (strata) of SNPs with increasing

significance of p-values in either ADHD or EA (formal definition of SNP stratum is given in supplementary material, available online). To further support this hypothesis, we estimated genetic correlation between ADHD and EA using LD score regression[8]. Details of these analyses are described in supplementary material.

To identify specific loci associated with ADHD, we applied the condFDR method described previously[9]. The condFDR method takes summary statistics that reflect genetic association of a phenotype of interest (primary) together with those of an auxiliary (conditional) phenotype and estimates a posterior probability that a SNP is null (has no association) in the primary phenotype, given that p-values of the SNP in both the primary and conditional phenotypes are lower than observed p-values. Thus, the condFDR method increases the power to discover loci associated with a primary phenotype by leveraging associations with a secondary phenotype. It does so by re-ranking SNPs compared to nominal p-value-based ranking[9]. In contrast, ranking SNPs based on unconditional FDR (e.g. using Benjamini–Hochberg or Benjamini–Yekutieli procedure) does not change their order (compared to nominal p-values).

Although both conditional QQ plots and genetic correlation based on the LD score regression can be useful to get a general idea of whether two traits have a significant genetic overlap, they are unable to find specific susceptibility loci shared by the traits. The conjFDR approach is an extension of condFDR allowing the identification of specific loci associated with both traits[9]. The conjFDR is defined as the maximum of the two condFDR values (taking one phenotype as primary and another as conditional and vice versa) for a specific SNP. Thus, the conjFDR approach estimates a posterior probability that an SNP is null for either phenotype or both at the same time, given that the p-values for both phenotypes are lower than the observed p-values. The method, therefore, uncovers loci associated with both phenotypes simultaneously.

To avoid inflation of the results due to LD-dependency in fold-enrichment and QQ plots as well as in condFDR/conjFDR analyses, we randomly pruned all SNPs across 500 iterations. For each iteration, all but one random SNP in each LD-independent region (clump of SNPs in strong LD, $r^2 > 0.2$) were removed, and finally the results were averaged across all iterations. LD (r^2 values) was estimated based on the 1000 Genomes Project phase 3 European sub-population data using PLINK[22].

As for meta-analyses based on multiple data-sources, the quality of our condFDR/conjFDR analysis will depend on the robustness of the primary data. More details about condFDR and conjFDR methods can be found in supplementary material and in the original publication[9].

Evaluation of the detected ADHD loci in an independent study of ADHD symptoms

We used genetic data on association of ADHD symptoms obtained from EAGLE consortium to test whether our results can be supported by data from the independent sample. For this purpose, we checked whether effects of the most significant SNPs in the loci identified by condFDR/conjFDR analyses are consistent between PGC ADHD and EAGLE data sets.

In silico identification of allele-specific effects of significant SNPs on transcription

Identifying and investigating genetic variants that might affect gene expression (expression quantitative trait loci or eQTLs) may shed light on how associated variants may contribute to biological mechanisms underlying a phenotype. eQTLs vary significantly both between different tissues and over time[23]. Existing GWASs on ADHD and EA clearly demonstrate remarkable enrichment of association signals in genomic regions implicated in regulation of gene expression in brain[18,20]. Hence, we focused on eQTL analysis of genes expressed in brain tissues. Significant associations identified with condFDR and conjFDR analyses were queried for known eQTLs using the GTEx portal (http://gtexportal.org) and the Braineac database (http://www.braineac.org). The latter database contains information on cis-eQTLs for 10 brain regions: cerebellar cortex, frontal

7

cortex, hippocampus, medulla (specifically inferior olivary nucleus), occipital cortex (specifically primary visual cortex), putamen, substantia nigra, thalamus, temporal cortex and intralobular white matter. Additionally, we checked age-dependent variations of expression in genes containing identified significant SNPs using the Human Brain Transcriptome database (http://hbatlas.org)[24].

Results

Evaluation of genetic overlap and correlation

In the absence of genetic overlap between two traits, it is expected that p-values for association with one trait are independent from the p-values for association with the other. However, conditional QQ plots in Figure 1 clearly demonstrate an increasing degree of leftward deflection for strata of more significant SNPs. This is observed both when conditioning ADHD on EA (Figure 1A) and vice versa (Figure 1B), suggesting substantial cross-trait polygenic enrichment. Enrichment of association signals for one trait among those of another is also clearly visible in the fold-enrichment plots, with more than 10-fold enrichment of SNPs from the strictest stratum ($p_{\text{conditional trait}} < 1.00 \times 10^{-3}$) for both traits (Figure S1, available online). Additionally, association z-scores of ADHD and EA demonstrate increasing negative correlation in more strictly defined strata of SNPs, both when strata are defined based on ADHD p-values (Figure 1C) and on EA p-values (Figure 1D). Moreover, LD score regression analysis also showed significant negative genetic correlation ($r_g = -0.403$, se = 0.075, $p = 7.90 \times 10^{-8}$) between these phenotypes.

Identification of ADHD-associated loci and loci shared between ADHD and EA

Using the condFDR/conjFDR method we identified 5 LD-independent regions, significantly associated with ADHD (condFDR < 0.01, conjFDR < 0.05), 3 of which were also identified as

shared between ADHD and EA. From each of these regions a single SNP with the lowest condFDR/conjFDR value (strongest association signal) was selected to represent their loci. These SNPs are presented in Table 1. Manhattan plots resulting from condFDR and conjFDR analyses are presented in Figures 2 and 3, respectively. Four out of five identified most significant SNPs revealed the opposite directions of effect in ADHD and EA.

Identified loci and related genes

Two loci (represented in Table 1 by variants rs618678 and rs412458) were identified both in condFDR and conjFDR analyses. rs618678 represents the strongest signal in the conjFDR analysis (conjFDR = 3.82×10^{-3}) and the second strongest in the condFDR analysis (condFDR = 3.77×10^{-3}). This SNP is an intronic variant within *KDM4A* on chromosome 1p34.2 (Figure 4B). Figure 4B and Figure S2B (available online) show the genetic context of rs618678, indicating, respectively, the conjFDR and condFDR values of adjacent SNPs. It is worth noting that in our analysis rs618678 tags a broad region of association. As can be seen in Figure 4B, multiple significant SNPs in strong LD ($r^2 >$ 0.60) with rs618678 were detected in this region, spanning over more than 200000 basepairs (bp). Besides *KDM4A*, the region also contains *PTPRF* (located in 1p34.2, upstream of *KDM4A*) and *ST3GAL3* (1p34.1, directly downstream *KDM4A*) genes. The latter was also identified in the eQTL analysis (discussed below). Another significant signal identified in both condFDR (condFDR = 7.34×10^{-3}) and conjFDR (conjFDR = 2.11×10^{-2}) analyses is represented by rs412458, an intronic variant within *MEF2C* on chromosome 5q14.3 (Figure S2A, D, available online).

Two loci were identified by condFDR, but not conjFDR. The strongest signal was detected at rs4324303 (condFDR = 2.17×10^{-3}), that is in the intergenic region on chromosome 2p24 (Figure 4A). Multiple significant variants tagged by rs4477079 (condFDR = 4.37×10^{-3}) were also identified on chromosome 8 within *RUNX1T1* (Figure S2C, available online).

9

Finally, conjFDR analysis identified a shared variant (conjFDR $= 4.48 \times 10^{-2}$) at *PINK1* (rs17414302, intronic, 1p36.12) (Figure S2E, available online). There were no LD-linked SNPs in the direct vicinity and only 25 SNPs in LD ($r^2 > 0.20$) with this variant, residing upstream of *PINK1,* at about 100000 bp.

None of SNPs identified either in condFDR or conjFDR reached genome-wide significance in previously published GWAS of ADHD[18]. Rs618678 reached genome-wide significance in EA (p $= 1.05 \times 10^{-10}$)[20]. Rs412458, which was identified by both condFDR and conjFDR, was not reported as genome-wide significant by the published EA GWAS (p $= 3.73 \times 10^{-6}$), but it is in LD ($r^2 = 0.35$) with rs588282 that did reach genome-wide significance in that study (previously reported p $= 1.69 \times 10^{-10}$). Other loci identified in our analyses were below genome-wide significance threshold in EA. It is also worth noting that the unconditional FDR values for all identified SNPs were above 0.01 and 0.05 in condFDR and conjFDR analysis respectively.

Evaluation of the detected ADHD loci in an independent study of ADHD symptoms

To assess the robustness of our results, we examined the loci identified in either the condFDR or conjFDR analyses (Table 1) in the association summary statistics from the independent GWAS of ADHD symptoms conducted by EAGLE consortium[21]. Four out of five loci (represented by SNPs: rs17414302, rs412458, rs618678, rs4324303) have the same direction of effect in the PGC and EAGLE GWASs while the last locus (represented by rs4477079 SNP) has an opposite direction of effect in these GWASs. These results are presented in Table S2 (available online).

In silico identification of allele-specific effects on transcription

According to Human Brain Transcriptome data[24], all six implicated genes (Table 1, Genes in the region) have a pronounced expression in different brain regions during the whole life cycle (Figure

S3, available online). Therefore, alterations in the expression level of these genes (where the detected SNPs are located) may affect a broad variety of processes over an extended period. We scanned the Braineac database to check whether SNPs identified in either the condFDR or conjFDR analyses are associated with gene expression in brain tissues. We found that four of five SNPs from Table 1 may operate as eQTLs, significantly (p < 0.001) associated with the expression of 13 different genes in several brain regions (Table S1, available online). Among those 13 genes, the most significant eQTL was observed between rs618678 and *ST3GAL3*. Further, significant eQTL effects of rs618678 on *ST3GAL3* were identified in muscle-skeletal tissue (p $= 3.40 \times 10^{-5}$) in the GTEx database (https://gtexportal.org/), but not in the brain tissue.

Discussion

The present study sought to investigate the genetic overlap between ADHD and EA, to leverage their potentially common genetics in order to improve the discovery of ADHD-associated loci and help our understanding of the correlation between EA and ADHD observed in epidemiological studies. It is, however, worth emphasizing the broad potential of the applied methodology, which can be used to leverage the great variety of existing GWAS data for dissecting the molecular genetic basis underlying complex human traits and disorders and their shared genetic etiology.

We identified significant genetic overlap between ADHD and EA supported by a pronounced genetic correlation ($r_g = -0.403$, se $= 0.075$, p $= 7.90 \times 10^{-8}$), consistent enrichment of shared variants in conditional QQ plots (Figure 1A, B), more than 10-fold mutual enrichment of SNPs associated with both traits (Figure S1, available online) and growing negative correlation of association z-scores for the nested SNP strata with increasing significance in both traits (Figure 1C, D). These findings

encourage the hypothesis that there is a shared genetic basis underlying ADHD and EA where in general ADHD risk alleles are associated with lower EA.

In comparison to previous study, exploring the topic of genetic overlap between ADHD and EA[17], our analysis employs a much larger data set of EA, allowing for a more reliable detection of genetic overlap (Figure 1; $r_g = -0.403$, se $= 0.075$, p $= 7.90 \times 10^{-8}$). It is also worth noting that we report a genetic correlation that is stronger than previously observed using the same ADHD data and a smaller (N $= 101069$) EA dataset ($r_g = -0.305$, se $= 0.141$, p $= 3.00 \times 10^{-2}$)[17]. Moreover, our study provides further insights into the shared genetic basis of ADHD and EA by identifying specific genetic loci jointly influencing these phenotypes. Further studies are warranted to determine in what way these genetic variants influence ADHD and EA. It is feasible that the shared genetic effects may influence EA through an intermediary phenotype such as reading disability, which is comorbid to ADHD[25], or through more basic neurobiological systems.

By combining GWAS summary statistics data on ADHD and EA[18,20] in the condFDR/conjFDR analyses, we enhanced discovery in the moderately powered ADHD GWAS and found five novel LD-independent loci associated with ADHD (Table 1). None of the loci identified in our analyses reached genome-wide significance in the ADHD GWAS[18], while rs618678 and rs412458 reached genome-wide significance in the GWAS of EA[20]. Four of five loci have opposite directions of effect in PGC case-control ADHD study[18] and EA study[20] (Table 1) and consistent directions of effect in the independent population-based study of ADHD symptoms from the EAGLE consortium[21] (Table S2, available online). The only SNP (rs4477079) having the same direction of effect in PGC ADHD data set and EA also has inconsistent effect directions in the PGC ADHD and EAGLE ADHD datasets. Despite the relatively small GWAS sample sizes on ADHD by the PGC[18] and EAGLE[21] consortia, and their differences in definitions of phenotype, observed consistency of effect directions of the

identified variants supports the credibility of the findings and the statistical approach. The fact that the majority of identified SNPs had opposite directions of effect in ADHD and EA is in line with the observed negative genetic correlation and corresponds to the expectations that can be drawn from existing clinical studies demonstrating poor academic performance and decreased rates of high school graduation and postsecondary education in individuals with diagnosed ADHD[14]. Altogether, these findings provide new insights into the genetic architecture of ADHD, suggesting shared molecular genetic mechanisms with EA. Furthermore, the findings may suggest that individuals with a high load of ADHD genetic risk factors, but not necessarily with the disorder itself, may be at higher risk for lower EA.

The most significant locus shared between ADHD and EA (rs618678) is located on chromosome 1 and represents a broad region of association spanning over more than 200,000 bp in 1p34.2 and 1p34.1 (Figure 4B; Figure S2B, available online). This region contains three protein coding genes: *PTPRF*, *KDM4A* and *ST3GAL3*. rs618678 is an intronic variant within *KDM4A*, a member of the Jumonji domain 2 family, which encodes a protein that demethylates histone residues, and acts as an epigenetic transcriptional regulator[26]. Genome-wide significant variants within *KDM4A* were reported in a recent GWAS of schizophrenia[5], a disorder that may share genetic background with ADHD. The protein encoded by *PTPRF* is a member of the protein tyrosine phosphatase (PTP) family, which regulates a variety of cellular processes, including cell growth, differentiation, mitotic cycle and oncogenic transformation. Mouse studies showed that *PTPRF* promotes neurogenesis in the hippocampus[27], a brain region linked to memory. *ST3GAL3* encodes a sialyltransferase responsible for the terminal sialylation of brain gangliosides and glycoproteins, which constitute a major part of the surface glycan coat of neurons and glia and act as an interface for cellular interactions[28]. Interestingly, mutations of *ST3GAL3* may impair the development of higher cognitive functions[29] and are associated with severe infantile epilepsy[30]. Our eQTL analysis with Braineac database revealed strong associations of

13

rs618678 with altered expression of *ST3GAL3* (Table S1, available online), suggesting that this may be a potential mechanism whereby this locus affects ADHD and EA. However, this association was not detected using GTEx database. The discrepancy between the results from the different eQTL-datasets could be attributed to differences in methodological techniques or sample configuration between the eQTL databases, or reflect the relatively small sample sizes. The eQTL results should be re-assessed when larger brain-eQTL databases are available.

The second locus shared between ADHD and EA (rs412458) is an intronic variant within *MEF2C* (Figure S2A, D, available online) which has multiple LD-linked variants with low condFDR/conjFDR values. *MEF2C* encodes one of four transcription factors constituting the myocyte enhancer factor 2 (MEF2) family[31]. MEF2 is involved in neuronal survival and may regulate the growth and pruning of neurons as well as the number of synapses in the hippocampus, with potential relevance for memory and learning[32]. Mutations of *MEF2C* cause severe mental retardation with stereotypic movements, seizures and/or cerebral malformations[33]. Further, genome-wide significant SNPs within *MEF2C* have been reported to be associated with schizophrenia[5] which shares polygenic risk with ADHD[34]. In addition, mutations in MEF2 genes have been found in patients with different neurological disorders including Rett-like disorder and Parkinson's diseases[32]. *MEF2C* expression is particularly enriched in the cerebral cortex[35] (Figure S3, available online).

The third locus identified as susceptible for both ADHD and EA by conjFDR is an intronic variant within *PINK1* on chromosome 1 (rs17414302). *PINK1* encodes a serine/threonine protein kinase that primarily localizes to mitochondria and protects against progressive mitochondrial damage and dysfunction[36]. This protein is thought to be involved in regulating neurite morphogenesis, enhancing anterograde mitochondrial transport and density of mitochondria in dendrites and upregulating expression of neuronal differentiation proteins[37]. *PINK1* is important for the maintenance of mitochondria in part by selective degradation of compromised mitochondria (mitophagy)[38]. Mutations

in this gene are a common cause of autosomal recessive Parkinson's disease[39]. However, rs17414302 represents an isolated signal with rather poor LD support (Figure S2E, available online) and it should thus be examined in more detail.

The strongest SNP association with ADHD revealed by the condFDR analysis was rs4324303. This SNP was not significant in the conjFDR analysis, but showed consistent direction of effect with ADHD symptoms in the EAGLE sample, possibly suggesting a putative role specific to ADHD. Rs4324303 is an intergenic variant located approximately 1 mega base upstream of the nearest protein coding gene (TRIB2). It is therefore difficult to speculate about the potential role of this variant in different cellular processes.

Another variant identified by the condFDR analysis is rs4477079, an intronic variant within *RUNX1T1* on chromosome 2. RUNX1T1 acts as a co-repressor of Notch[40] and Wnt[41] pathways. *RUNX1T1* was reported to have high expression levels in adult and fetal brain[42] and may influence axon guidance process[43]. *RUNX1T1* was previously identified among the top associations (although not reaching genome-wide significance) in the context of oppositional defiant disorder (ODD), which is a frequent psychiatric disorder seen in individuals with ADHD[44]. Notably, unlike the other loci identified in our analyses, this locus shows an inconsistent direction of effect between PGC ADHD risk and quantitative measures of ADHD symptoms in pediatric populations (Table S2) and a co-directional effect between PGC ADHD risk and EA (Table 1). The latter is contrary to expectations based on previous findings. The role of *RUNX1T*, thus, remains puzzling and further studies are needed to clarify it.

To further evaluate the ADHD-associated variants identified in this study utilizing the data from PGC ADHD case-control and EA GWASs, we examined our top hits in the light of the ADHD symptoms' GWAS. Four of five loci identified here revealed consistent direction of effect in the independent

GWAS of ADHD symptoms (Table S2, available online). Of note, twin studies provide strong evidence that the diagnosis of ADHD can be considered as the extreme of a continuous trait[45] and several studies show that the polygenic risk score computed from an association study of ADHD diagnosis predicts the variability of ADHD symptoms in population samples[21,46]. Additionally, it has been shown that the continuous measure of ADHD (such as symptom score) and the ADHD diagnosis share over 90% of their genetic background[47]. Thus, the results of the performed exploration may be viewed as confirmatory of our findings.

It is also worth mentioning that two loci identified in our analyses (corresponding to rs618678 and rs412458 in Table 1) were reported to reach genome-wide significance in the largest GWAS on ADHD performed to date, with the total number of 20183 ADHD cases and 35191 controls. In this GWAS, ADHD diagnosis was based on either ICD10 or DSM-IV. The study is yet unpublished but preprint is available in bioRxiv[47].

As children with ADHD have been reported to have high risk for academic failure, school dropout, grade repetition and placement in special education[48,49], it is likely that the prevalence of ADHD cases among individuals with lower EA would be increased compared to the prevalence among individuals with higher EA. Moreover, ADHD is known to have a complex pattern of co-morbid conditions[1] (including dyslexia[25], oppositional defiant disorder[50] and others), many of them are also associated with lower EA. This potential overlap of phenotypes prevents us from translating the genetic correlation into actual pleiotropy, which is defined as the same gene variant affecting independent diseases or traits. Furthermore, it is challenging to evaluate small effect sizes, and speculate about molecular mechanisms behind the effective variants when examining such potentially overlapping phenotypes. Another general problem is that the effects of the associated variants are small and their functional roles have not been directly investigated. Associated genetic loci contain several genes and it is difficult to establish an arrow of causality when studying association between traits. Thus, the

16

question whether ADHD is diagnosed because of observed educational problems or ADHD is the cause of subsequent educational problems or there is other common underlying factor needs further exploration.

Also of possible relevance is the sample overlap between PGC ADHD and EA datasets (both GWASs include the WTCCC58C cohort[51]), which may inflate the results of our FDR analyses. However, the results of LD score regression, which are in line with those of our FDR analyses, are not affected by the sample overlap[8].

We identified five loci associated with ADHD and provided evidence for a shared genetic basis between ADHD and EA, implicating three genetic loci in this overlap. Four of five identified loci showed consistent effects in the independent data set of ADHD symptoms, and inverse correlation with EA, in line with prior epidemiological and genetic studies. Altogether, the findings provide new insights into the relationship between ADHD and EA, suggesting shared molecular genetic mechanisms. On a cautious note, the identified risk variants are not informative clinically due to their small effect sizes. Further research is required to clarify the biological effects of the identified genetic variants and how these may influence EA and ADHD pathogenesis.

References

1. Faraone SV, Asherson P, Banaschewski T, et al. Attention-deficit/hyperactivity disorder. *Nat Rev Dis Primers.* 2015;1:15020.
2. Cross-Disorder Group of the Psychiatric Genomics C, Lee SH, Ripke S, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet.* 2013;45(9):984-994.
3. Freitag CM, Rohde LA, Lempp T, Romanos M. Phenotypic and measurement influences on heritability estimates in childhood ADHD. *Eur Child Adolesc Psychiatry.* 2010;19(3):311-323.
4. Martin J, O'Donovan MC, Thapar A, Langley K, Williams N. The relative contribution of common and rare genetic variants to ADHD. *Transl Psychiatry.* 2015;5:e506.
5. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421-427.

6. Bipolar Disorder Working Group PGC. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet.* 2011;43(10):977-983.

7. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012;90(1):7-24.

8. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-1241.

9. Andreassen OA, Djurovic S, Thompson WK, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet.* 2013;92(2):197-209.

10. Desikan RS, Schork AJ, Wang Y, et al. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. *Circulation.* 2015;131(23):2061-2069.

11. Andreassen OA, Thompson WK, Dale AM. Boosting the power of schizophrenia genetics by leveraging new statistical tools. *Schizophr Bull.* 2014;40(1):13-17.

12. Ferrari R, Wang Y, Vandrovcova J, et al. Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. *J Neurol Neurosurg Psychiatry.* 2017;88(2):152-164.

13. Yokoyama JS, Wang Y, Schork AJ, et al. Association Between Genetic Traits for Immune-Mediated Diseases and Alzheimer Disease. *JAMA Neurol.* 2016;73(6):691-697.

14. Voigt RG, Katusic SK, Colligan RC, Killian JM, Weaver AL, Barbaresi WJ. Academic Achievement in Adults with a History of Childhood Attention-Deficit/Hyperactivity Disorder: A Population-Based Prospective Study. *J Dev Behav Pediatr.* 2017;38(1):1-11.

15. Barkley RA. Major life activity and health outcomes associated with attention-deficit/hyperactivity disorder. *J Clin Psychiatry.* 2002;63 Suppl 12:10-15.

16. Biederman J, Newcorn J, Sprich S. Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *Am J Psychiatry.* 1991;148(5):564-577.

17. Hagenaars SP, Harris SE, Davies G, et al. Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Mol Psychiatry.* 2016;21(11):1624-1632.

18. Neale BM, Medland SE, Ripke S, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry.* 2010;49(9):884-897.

19. Zayats T, Jacobsen KK, Kleppe R, et al. Exome chip analyses in adult attention deficit hyperactivity disorder. *Transl Psychiatry.* 2016;6(10):e923.

20. Okbay A, Beauchamp JP, Fontana MA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature.* 2016;533(7604):539-542.

21. Middeldorp CM, Hammerschlag AR, Ouwens KG, et al. A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry.* 2016;55(10):896-905 e896.

22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575.

23. Francesconi M, Lehner B. The effects of genetic variation on gene expression dynamics during development. *Nature.* 2014;505(7482):208-211.

24. Kang HJ, Kawasawa YI, Cheng F, et al. Spatio-temporal transcriptome of the human brain. *Nature.* 2011;478(7370):483-489.

25. Germano E, Gagliano A, Curatolo P. Comorbidity of ADHD and dyslexia. *Dev Neuropsychol.* 2010;35(5):475-493.

26. Guerra-Calderas L, Gonzalez-Barrios R, Herrera LA, Cantu de Leon D, Soto-Reyes E. The role of the histone demethylase KDM4A in cancer. *Cancer Genet.* 2015;208(5):215-224.

27. Bernabeu R, Yang T, Xie Y, Mehta B, Ma SY, Longo FM. Downregulation of the LAR protein tyrosine phosphatase receptor is associated with increased dentate gyrus neurogenesis and an increased number of granule cell layer neurons. *Mol Cell Neurosci.* 2006;31(4):723-738.

28. Yoo SW, Motari MG, Susuki K, et al. Sialylation regulates brain structure and function. *FASEB J.* 2015;29(7):3040-3053.

29. Hu H, Eggers K, Chen W, et al. ST3GAL3 mutations impair the development of higher cognitive functions. *Am J Hum Genet.* 2011;89(3):407-414.

30. Edvardson S, Baumann AM, Muhlenhoff M, et al. West syndrome caused by ST3Gal-III deficiency. *Epilepsia.* 2013;54(2):e24-27.

31. Potthoff MJ, Olson EN. MEF2: a central regulator of diverse developmental programs. *Development.* 2007;134(23):4131-4140.

32. Dietrich JB. The MEF2 family and the brain: from molecules to memory. *Cell Tissue Res.* 2013;352(2):179-190.

33. Le Meur N, Holder-Espinasse M, Jaillard S, et al. MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J Med Genet.* 2010;47(1):22-29.

34. Nivard MG, Gage SH, Hottenga JJ, et al. Genetic Overlap Between Schizophrenia and Developmental Psychopathology: Longitudinal and Multivariate Polygenic Risk Prediction of Common Psychiatric Traits During Development. *[published online ahead of print Mar 2017] Schizophr Bull.* DOI: 10.1093/schbul/sbx031.

35. Lin X, Shah S, Bulleit RF. The expression of MEF2 genes is implicated in CNS neuronal differentiation. *Brain Res Mol Brain Res.* 1996;42(2):307-316.

36. Narendra DP, Jin SM, Tanaka A, et al. PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. *PLoS Biol.* 2010;8(1):e1000298.

37. Dagda RK, Pien I, Wang R, et al. Beyond the mitochondrion: cytosolic PINK1 remodels dendrites through protein kinase A. *J Neurochem.* 2014;128(6):864-877.

38. Scarffe LA, Stevens DA, Dawson VL, Dawson TM. Parkin and PINK1: much more than mitophagy. *Trends Neurosci.* 2014;37(6):315-324.

39. Valente EM, Abou-Sleiman PM, Caputo V, et al. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science.* 2004;304(5674):1158-1160.

40. Salat D, Liefke R, Wiedenmann J, Borggrefe T, Oswald F. ETO, but not leukemogenic fusion protein AML1/ETO, augments RBP-Jkappa/SHARP-mediated repression of notch target genes. *Mol Cell Biol.* 2008;28(10):3502-3512.

41. Moore AC, Amann JM, Williams CS, et al. Myeloid translocation gene family members associate with T-cell factors (TCFs) and influence TCF-dependent transcription. *Mol Cell Biol.* 2008;28(3):977-987.

42. Wolford JK, Prochazka M. Structure and expression of the human MTG8/ETO gene. *Gene.* 1998;212(1):103-109.

43. Terman JR, Kolodkin AL. Nervy links protein kinase a to plexin-mediated semaphorin repulsion. *Science.* 2004;303(5661):1204-1207.

44. Aebi M, van Donkelaar MM, Poelmans G, et al. Gene-set and multivariate genome-wide association analysis of oppositional defiant behavior subtypes in attention-deficit/hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet.* 2016;171(5):573-588.

45. Levy F, Hay DA, McStephen M, Wood C, Waldman I. Attention-deficit hyperactivity disorder: a category or a continuum? Genetic analysis of a large-scale twin study. *J Am Acad Child Adolesc Psychiatry.* 1997;36(6):737-744.

46. Groen-Blokhuis MM, Middeldorp CM, Kan KJ, et al. Attention-deficit/hyperactivity disorder polygenic risk scores predict attention problems in a population-based sample of children. *J Am Acad Child Adolesc Psychiatry.* 2014;53(10):1123-1129 e1126.

47. Demontis D, Walters RK, Martin J, et al. Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv.* 2017.

48. Heiligenstein E, Guenther G, Levy A, Savino F, Fulwiler J. Psychological and academic functioning in college students with attention deficit hyperactivity disorder. *J Am Coll Health.* 1999;47(4):181-185.

49. Gray SA, Fettes P, Woltering S, Mawjee K, Tannock R. Symptom Manifestation and Impairments in College Students With ADHD. *J Learn Disabil.* 2016;49(6):616-630.

50. Wilens TE, Biederman J, Brown S, et al. Psychiatric comorbidity and functioning in clinically referred preschool children and school-age youths with ADHD. *J Am Acad Child Adolesc Psychiatry.* 2002;41(3):262-268.

51. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol.* 2006;35(1):34-41.

52. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336-2337.

Tables

SNP	Chr Region	Position	condFDR	conjFDR	Location relative to gene	Genes in the region	p-value ADHD	p-value EA	Effect size ADHD	Effect size EA
rs17414302	1p36.12	20976535	4.45×10^{-2}	**4.48×10^{-2}**	intronic	***PINK1***	1.97×10^{-4}	8.17×10^{-7}	-0.090	0.022
rs618678	1p34.2	44133299	**3.77×10^{-3}**	**3.82×10^{-3}**	intronic	***KDM4A*** *PTPRF* *ST3GAL3*	1.05×10^{-5}	1.05×10^{-10}	-0.053	0.017
rs4324303	2p24	13817678	**2.17×10^{-3}**	5.75×10^{-1}	intergenic	-	2.05×10^{-7}	6.69×10^{-3}	-0.079	0.009
rs412458	5q14.3	88029627	**7.34×10^{-3}**	**2.11×10^{-2}**	intronic	***MEF2C***	2.15×10^{-5}	3.73×10^{-6}	0.061	-0.014
rs4477079	8q21.3	93059038	**4.37×10^{-3}**	3.65×10^{-1}	intronic	***RUNX1T1***	1.44×10^{-6}	1.62×10^{-3}	-0.071	-0.009

Table 1. Most significant SNPs for each LD-independent region identified either with condFDR (**condFDR < 0.01**) or with conjFDR (**conjFDR < 0.05**) analysis. condFDR/conjFDR values that are below the predefined significance threshold of 0.01/0.05 are marked with bold. Chromosome and position are indicated according to GRCh37. For both ADHD and EA, p-values without genomic inflation correction are shown. Thee effect size is given as **$\log_{10}(OR)$** for ADHD and as Beta regression coefficient for EA. Genes in the region are defined as genes containing SNPs at either **condFDR < 0.01** or **conjFDR < 0.05** and in LD (**$r^2 > 0.20$**) with the most significant SNP of the locus. Genes containing the leading SNP are marked in bold. Annotation was generated with Biomart Variant Effect Predictor (http://www.ensembl.org/Homo_sapiens/Tools/VEP).

Figure legends

Figure 1. Conditional QQ plots and correlation plots.

Conditional QQ plots (A, B) demonstrate relation between expected (x axis) and observed (y axis) significance of markers in the primary trait when markers are stratified by their p-values in the conditional trait. A sequence of four nested strata is presented: all SNPs (i.e. p-values of the conditional trait ≤ 1.00), $p_{\text{conditional trait}} < 0.1$, $p_{\text{conditional trait}} < 0.01$ and $p_{\text{conditional trait}} < 0.001$.

A: ADHD conditioned on educational attainment (EA).

B: educational attainment (EA) conditioned on ADHD.

Correlation plots (C, D) show Pearson's correlation coefficients between association z-scores of ADHD and EA for the nested strata of SNPs (as introduced in the conditional QQ plots) averaged over 500 iterations of random pruning. Solid black lines indicate standard deviations.

C: SNP strata are defined by the p-values of markers in educational attainment (ADHD|EA).

D: SNP strata are defined by the p-values of markers in ADHD (EA|ADHD).

Figure 2. Manhattan of $-\log_{10}(\text{FDR})$ for ADHD conditional on educational attainment (EA).

The data are unpruned. The small points are non-significant SNPs, the bold points represent significant SNPs (condFDR < 0.01). Points corresponding to significant SNPs with lowest conditinoal FDR in each LD-independent region ($r^2 > 0.20$) have a black border and either the name of corresponding gene (for SNPs within the gene) or the rs-number (for an intergenic SNP) written above it. The horizontal grey dotted line shows the significance threshold of condFDR (0.01). Black dots stand for unconditional FDR values.

Figure 3. Manhattan plot of conjunctional $-\log_{10}(\text{FDR})$ for ADHD and educational attainment (EA).

The data are unpruned. The small points stand for non-significant SNPs, the bold points represent significant SNPs (conjFDR < 0.05). Points corresponding to significant SNPs with lowest

conjunctional FDR in each LD-independent region ($r^2 > 0.20$) have a black border and the name of the corresponding gene written above it. The horizontal grey dotted line shows the significance threshold of conjFDR (0.05).

Figure 4. Genetic context of the strongest associations identified in condFDR (A) and conjFDR (B) analyses.

Values for both genotyped and imputed variants are shown on the left y-axis as $-\log_{10}(\text{condFDR})$ and $-\log_{10}(\text{conjFDR})$ respectively. In each subplot, an SNP with the strongest association is shown in the large purple square. The color of the remaining markers reflects the degree of LD with the strongest associated SNP measured as r^2 coefficient (described in the legend). The recombination rate is plotted as a blue solid line, its value in centimorgan/megabase (cM/Mb) is indicated on the right y-axis. The red dotted lines indicate the FDR thresholds (0.01 for condFDR and 0.05 for conjFDR).

A: surrounding of the strongest association in condFDR analysis: rs4324303 ($\text{condFDR} = 2.17 \times 10^{-3}$).

B: surrounding of the strongest association in conjFDR analysis: rs618678 ($\text{conjFDR} = 3.82 \times 10^{-3}$).

Figures are generated with LocusZoom[52].
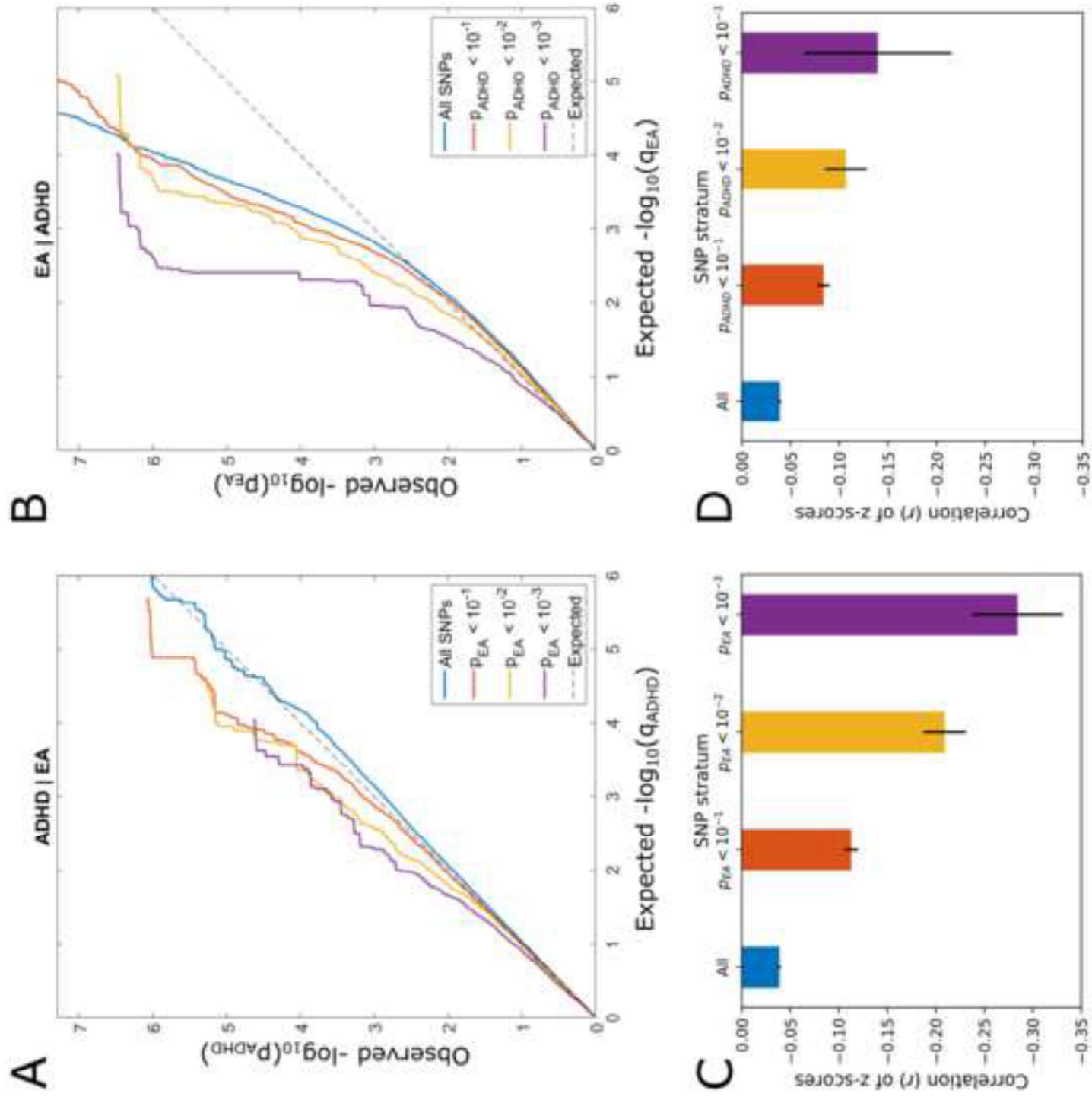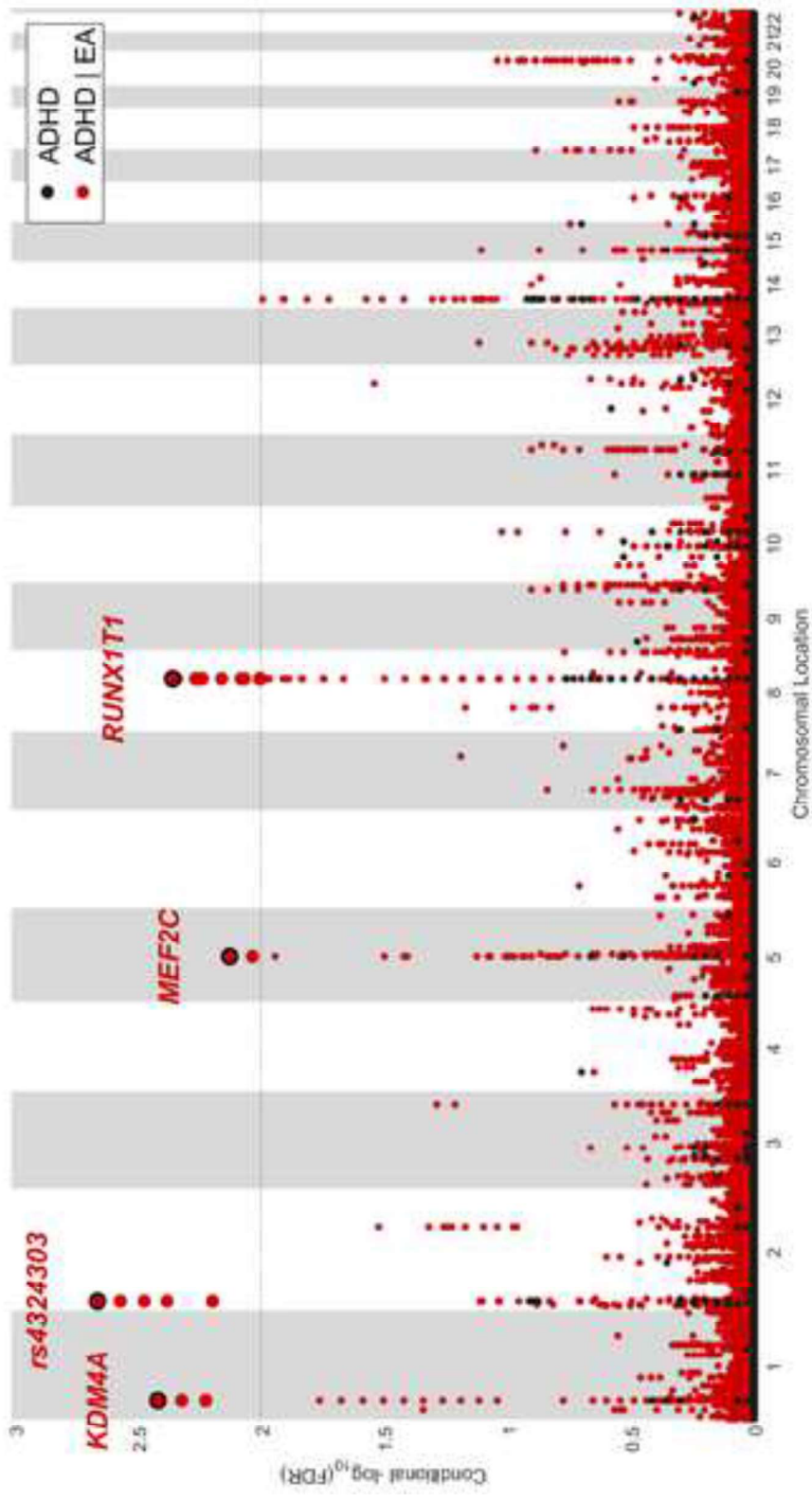
Figure 1
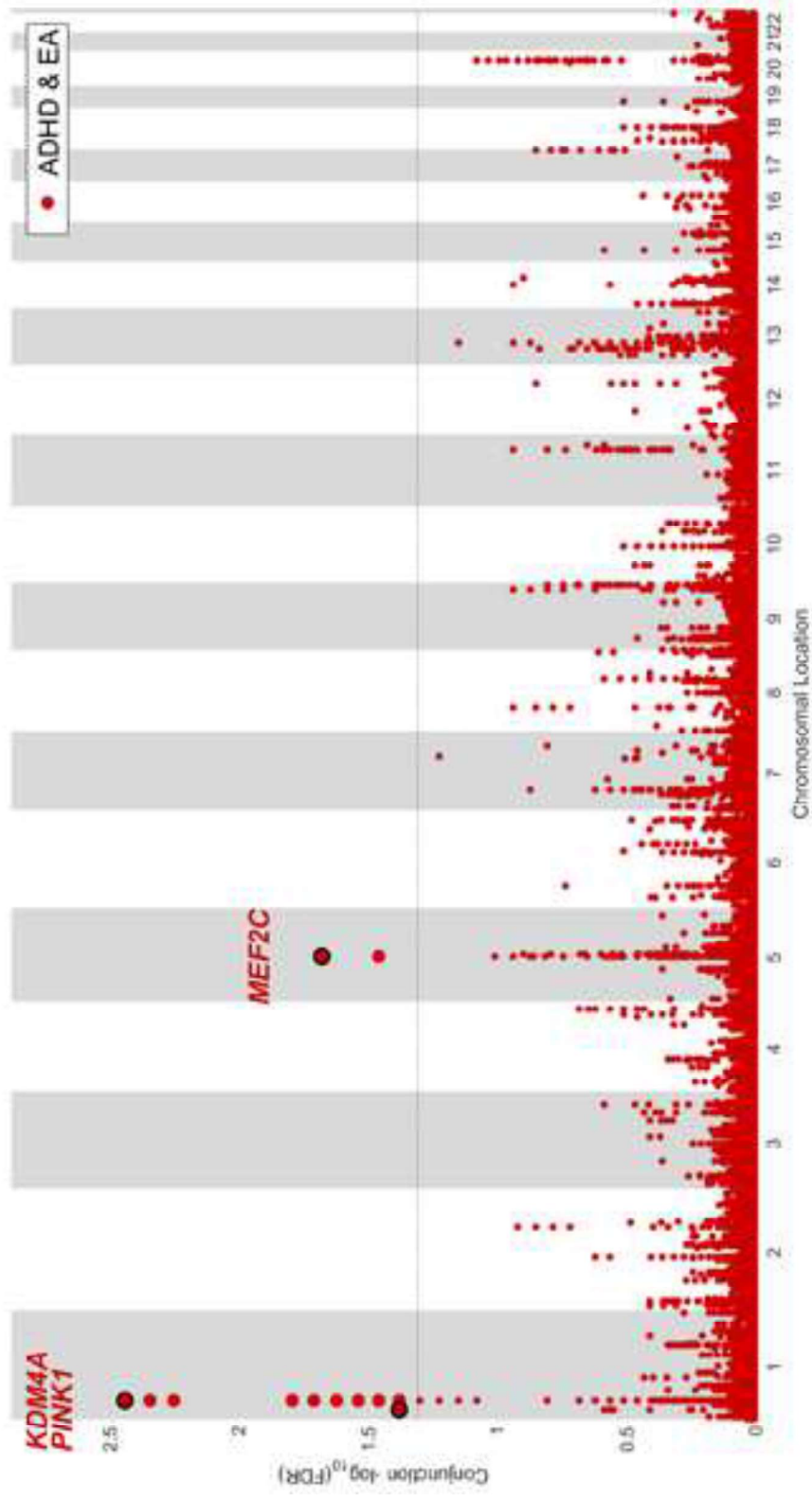
Figure 2

Figure 3

Figure 4

Supplemental material

Participant samples

We used ADHD data from the Psychiatric Genomics Consortium (PGC)[1]. The data set contains information from 2064 trios, 896 cases and 2455 controls combined from four independent studies (CHOP[2], IMAGE[3], IMAGE II[4], PUWMa[5]). Each dataset has undergone stringent quality control (including filtering out SNPs with high missingness and significant deviation from Hardy-Weinberg equilibrium) and was imputed using HapMap Phase III European CEU and TSI samples as the reference panels. The assoc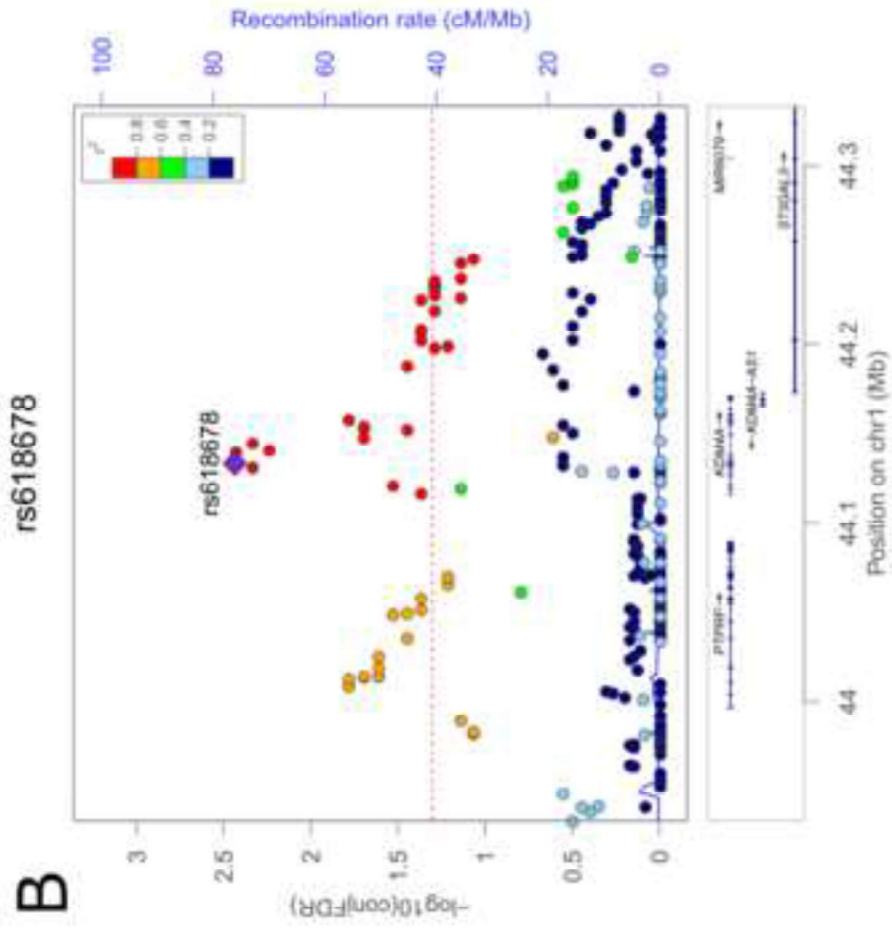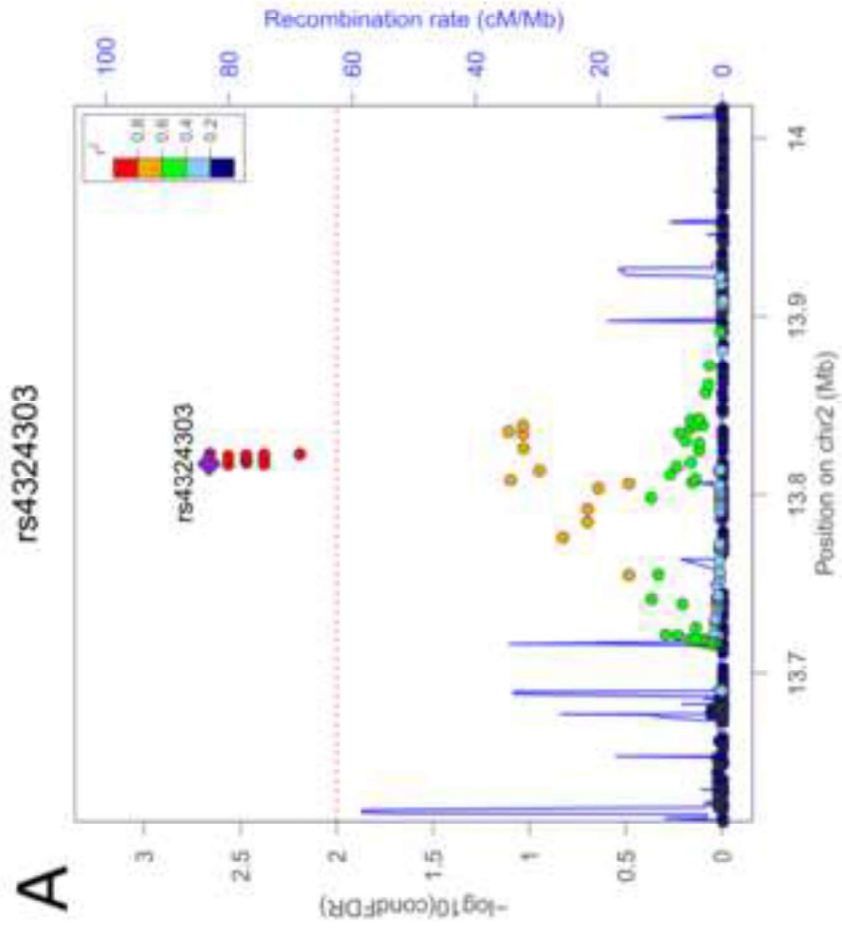iation analysis was performed separately on each dataset and the inverse variance weight meta-analysis has been carried out. Details of the data processing steps and the analyses are described in the original publication[1]. Prior to our analyses, we performed basic quality control of the obtained PGC summary statistics: SNPs with low imputation quality (info score < 0.80) and minor allele frequency (MAF) < 0.01 were excluded. Additionally, we removed the major histocompatibility complex (MHC) region (defined on hg19 as chr6: 28477797 - 33448354). The remaining 6393963 SNPs were used for the analyses. For quantile-quantile (QQ) and enrichment plots as well as for conjFDR analyses, all p-values were adjusted by genomic control inflation factor as described previously[6,7].

Different ADHD diagnostic procedures were used across studies. Below we briefly summarize phenotype measurement procedures in each study (taken from the original publications).

- CHOP: trio families were recruited from pediatric and behavioral health clinics in the Philadelphia area. Inclusion criteria included families of European descent with an ADHD proband (age 6 - 18). Exclusion criteria included prematurity (< 36 weeks), mental retardation, major medical and neurological disorders, pervasive developmental disorder, psychoses and major mood disorders. A child psychiatrist assessed diagnostic status of ADHD probands by K-SADS P-IVR interview. Parental ADHD was assessed using the ADHD Self-Report Scale.

- IMAGE: trio families of European origin were collected using a common protocol with centralized training and reliability testing of raters and centralized data management. Families were identified through ADHD probands aged 5 to 17 attending outpatient clinics at the data collection sites in Europe and Israel. Exclusion criteria were autism, epilepsy, IQ < 70, brain disorders and any genetic or medical disorder associated with externalizing behaviors that might mimic ADHD. Parents of children were interviewed with the Parental Account of Childhood Symptom (PACS), a semi-structured, standardized, investigator-based interview developed as an instrument to provide an objective measure of child behavior. Both parents and teachers completed the respective versions of the Conners ADHD rating scales and the Strengths and Difficulties Questionnaire. Using results of these surveys, probands had been clinically diagnosed as ADHD based on the DSM-IV criteria (or hyperkinetic disorder, the most closely equivalent category in the ICD-10 nomenclature used at some of the clinics).

- IMAGE II: a case-control study using samples collected by the IMAGE project (but not included in the IMAGE GWAS) and samples collected at additional sites (Germany, Scotland and Cardiff, UK) that were assessed in a manner similar to IMAGE samples. Cases were identified mainly through outpatient clinics at the data collection sites.

- PUWMa: trio families were collected independently at three sites (MGH, Washington University and UCLA) using similar but slightly different methods. Children were 5 - 19 years of age at initial assessment and met criteria for DSM-IV-TR ADHD.

  - MGH. Psychiatric assessments were made with K-SADSE (Epidemiologic Version) interview (indirect interviews with parents and additionally direct interviews with subjects older than 12). Exclusion criteria included major sensorimotor handicaps (deafness, blindness), psychosis/schizophrenia, autism, inadequate command of the English language, or a Full Scale IQ less than 80.

  - Washington University. Parents reported on their children and themselves, and the youths on themselves, using the Missouri Assessment of Genetics Interview for Children (MAGIC), a semi-structured psychiatric interview. DSM-IV diagnoses of ADHD were based upon parental reports. Families were excluded if a parent/guardian reported mental retardation or if the parent/guardian and twins could not speak English.

  - UCLA. Lifetime psychiatric diagnoses were based on semi-structured diagnostic interviews conducted by master's level clinical psychologists or highly trained interviewers with extensive experience and reliability training in psychiatric diagnoses. Children and adolescents were assessed using the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL). Adult parents were assessed using the Schedule for Affective Disorders and Schizophrenia-Lifetime version (SADS-LA-IV), supplemented with the K-SADS Behavioral Disorders module for diagnosis of ADHD. Direct interviews were supplemented with parent and teacher versions of the Swanson, Nolan, and Pelham, version IV (SNAP-IV) rating scale, as well as a parent-completed CBCL and Teacher Report Form. Parents also completed current ratings of self and spouse behavior with the ADHD Rating Scale IV. Best estimate diagnoses were assigned using all of the available clinical information according to strict DSM-IV criteria and reviewed by senior clinicians. Subjects were excluded from participation if they were positive for any of the following: neurological disorder, head injury resulting in concussion, lifetime diagnoses of schizophrenia or autism, or estimated Full Scale IQ < 70.

EA data were obtained from the Social Science Genetic Association Consortium (SSGAC)[8], where EA was derived from participants' number of years of education. As each country educates its population under various educational systems, the International Standard Classification of Education (ISCED) of the United Nations Educational, Scientific and Cultural Organization

(UNESCO) was used. Each major educational qualification that it is possible to attain in a specific country was mapped into one of seven ISCED categories. To construct the primary outcome variable, each ISCED category was then translated into US years of schooling. So eventually educational attainment for all subjects in the study (regardless of country of birth) was measured in US years of schooling.. For our analyses, we used summary statistics generated by the meta-analysis of all discovery and replication cohorts, except the 23andMe sample which is not publicly available (64 datasets with total N = 328917). For each dataset, genome-wide association was tested using only individuals of European descent, with EA assessed at age 30 or above. Cohort-level data underwent various quality control procedures (filtering out SNPs with high missingness and significant deviation from Hardy-Weinberg equilibrium etc.) and a round of genomic control. Sample-size-weighted meta-analysis of all 9256490 autosomal SNPs (from the 1000 Genomes Project) passing quality control procedures was performed using METAL[9]. Detailed information on sample selection, cohort-level quality control and meta-analysis can be found in the original publication[8]. Before conducting our analyses, we removed MHC region. In addition, for QQ plots, enrichment plots and conjFDR analyses, we also corrected p-values for genomic inflation (as defined above). The data are available for downloading at (http://www.thessgac.org/data).

We examined our top association signals in the summary statistics from an independent GWAS of ADHD symptoms performed by EArly Genetics and Lifecourse Epidemiology (EAGLE) consortium[10]. Unlike the PGC case-control ADHD GWAS, EAGLE GWAS represents a meta-analysis of 9 population-based pediatric cohorts containing information on 17666 children under the age of 13 years with measures of ADHD symptom scores. Several different measures of ADHD symptom scores were used across cohorts, including the Attention Problems scale of the Child Behavior Checklist (CBCL) and the Teacher Report Form (TRF), the Hyperactivity scale of the Strengths and Difficulties Questionnaire (SDQ) and the *DSM-IV* ADHD items (e.g. as included in the Conners' Rating Scale)[11-15]. For the meta-analysis, one phenotype was selected from each cohort. Based on the phenotype that was most available, school-age ratings were chosen over preschool-age ratings, parent ratings over teacher ratings, and the measurement instrument with the largest information density was preferred over the other instruments (Conners' DSM-IV > CBCL > SDQ). The study did not detect genome-wide significant SNPs.

LD score regression

To support a hypothesis of shared genetic basis between ADHD and EA we applied linkage disequilibrium (LD) score regression [16] to estimate SNP-based genetic correlation between them. For this calculation, we used only SNPs overlapping with HapMap phase 3 variants, while low quality imputed markers as well as indels, rare variants (MAF < 0.01) and markers from the MHC region were removed from both data sets as described in the main text. The analysis was performed using the Python-based package available at (https://github.com/bulik/ldsc) and the procedure is described in the documentation of the package (https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation). Since we know that there is sample overlap between ADHD and EA datasets, we didn't constrain regression intercept. Obtained genetic correlation was significantly

below zero (r_g = -0.403, SE = 0.075). Similar values can be obtained using LD Hub (http://ldsc.broadinstitute.org) with PGC ADHD data set and "years of schooling" data set available at LD Hub. The obtained genetic correlation is in line with our findings obtained with conjFDR model, where all three markers are significantly associated with both ADHD and EA presenting opposite directions of effect (Table 1). It is worth noting here that LD score based genetic correlation allows effective detection of genetic correlation only when the bulk of variants associated with both traits reveals strong net correlation between the direction of effects of overlapping SNPs relative to each other (i.e. correlated same or opposite direction of effects in two phenotypes, but not mixed). The method is also not able to identify specific shared loci. To circumvent the latter, we applied conjFDR method described below.

Fold enrichment plot

The fold enrichment plot allows the assessment of genetic enrichment in one (primary) trait when conditioning on another (conditional) trait. Enrichment is present if the degree of upward deflection from the expected null level (horizontal line through 1) depends on the stratum defined by the p-values for association with the trait used for conditioning (e.g. $p_{conditinal\ trait}$ < 1.00×10^{-1}).

Here by SNP stratum corresponding to some p-value threshold p_{thresh} in trait T (either ADHD or EA) we mean a set of SNPs for which z-scores are available in both ADHD and EA GWASs and which have p-values in trait T below the threshold p_{thresh}.

First, the empirical cumulative distribution function (CDF) of primary trait association p-values is computed for all SNPs. Then the CDF of association p-values is also estimated for each SNP stratum defined by the p-values for association with the conditional trait. Then fold enrichment for each stratum is estimated as the ratio $CDF_{stratum}/CDF_{all}$. The x-axis displays nominal $-\log_{10}(p\text{-value})$ for the primary trait, the y-axis shows the fold enrichment. Here we focus on polygenic effects for SNPs not reaching the standard GWAS significance threshold $-\log_{10}(p)$ < 7.30 (corresponding to $p > 5.00 \times 10^{-8}$).
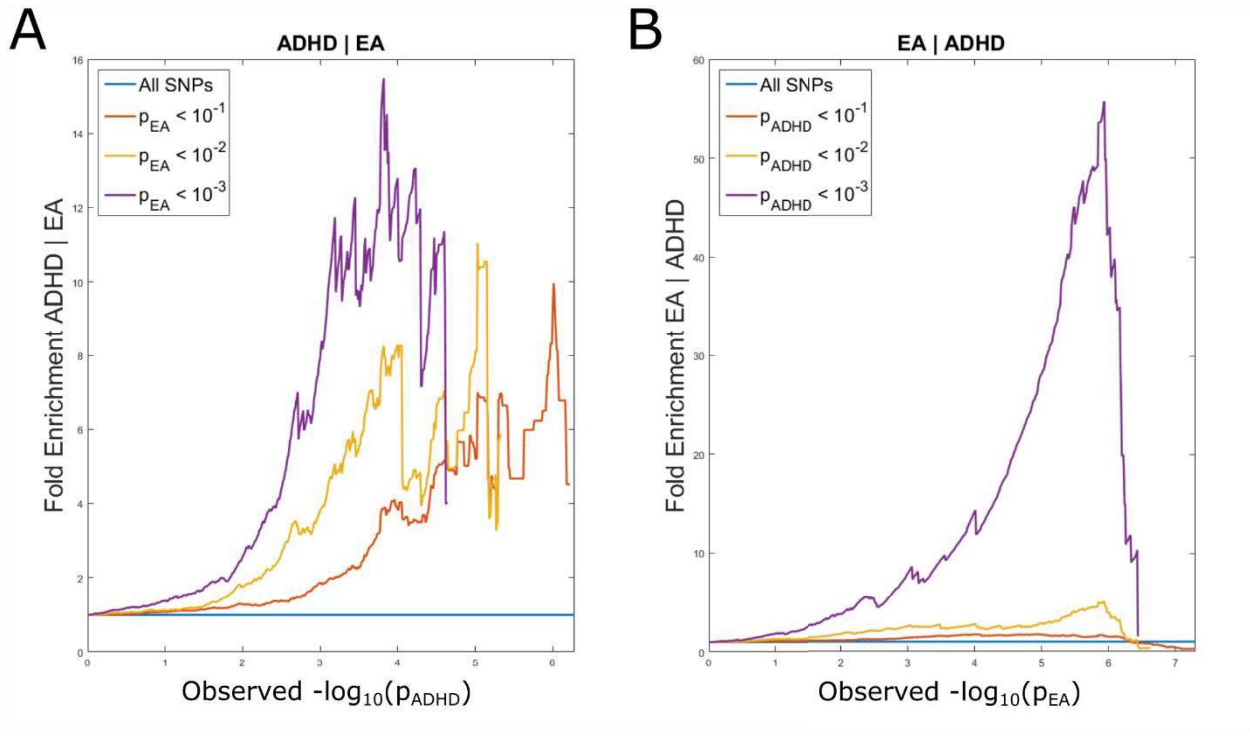
Figure S1. Fold enrichment of association between ADHD and educational attainment (EA). Fold enrichment plots of the observed $-\log_{10}(p)$ below the standard GWAS threshold (corresponding to $p > 5\times10^{-8}$) in the primary trait stratified based on the association with the conditional trait. A sequence of four nested strata is presented: all SNPs (i.e. p-values of the conditional trait <= 1), $p_{conditinal\ trait} < 10^{-1}$, $p_{conditionoal\ trait} < 10^{-2}$ and $p_{conditinoal\ trait} < 10^{-3}$. Successive upward elevation compared to all SNPs demonstrates polygenic enrichment both for ADHD conditioned on educational attainment (A) and for educational attainment conditioned on ADHD (B).

Conditional/Conjunctional FDR

The following brief description of conditional/conjunctional false discovery rate (condFDR/conjFDR) method is based on the papers [6,7] where the method was introduced and subsequent correction to the second paper [17].

To explain the condFDR/conjFDR method, we will first review the concept of standard (unconditional) false discovery rate (FDR). Then we describe the condFDR method. The latter is an extension of the standard FDR, which incorporates information from genome-wide association summary statistics of a second phenotype to adjust its significance level. After that we will present a formal definition of conjFDR for two phenotypes and show that it can be estimated as a maximum of two conditional FDRs.

In empirical Bayes interpretation, for a given p-value cutoff, FDR can be defined as follows [18]:

$$FDR(p) = \frac{\pi_0 F_0(p)}{F(p)}$$

where π_0 is the *a priori* fraction of null SNPs, F_0 is the null cumulative distribution function (CDF), and F is the CDF of all SNPs, both null and non-null. Under the null hypothesis, F_0 is the CDF of the uniform distribution on the unit interval [0,1], so $F_0(p) = p$ and the latter formula reduces to: $FDR(p) = \pi_0 p / F(p)$. Having this definition, the conditional FDR for two phenotypes can be defined as the posterior probability that a given SNP is null for the first phenotype given that the p-values for both phenotypes are as small as or smaller than the observed p-values. Formally this can be expressed as:

$$CondFDR(p_1, p_2) = \frac{\pi_0(p_2)p_1}{F(p_1|p_2)}$$

where p_1 and p_2 are p-values of SNPs in the first and the second phenotypes correspondingly, $\pi_0(p_2)$ is the conditional proportion of null SNPs and $F(p_1|p_2)$ is the conditional cdf for the first phenotype given that p-values for the second phenotype are p_2 or smaller. We denote the conditional FDR for phenotype 1 (pt1) given phenotype 2 (pt2) as $FDR_{pt1|pt2}$. In our calculations, we produce a conservative estimate of $FDR_{pt1|pt2}$ by setting $\pi_0(p_2) = 1$ and using the empirical conditional cdf in place of $F(p_1|p_2)$.

In our study, conjFDR is used to identify SNPs that are associated with two phenotypes simultaneously. It is defined as the posterior probability that a given SNP is null for either phenotype or both phenotypes simultaneously when the p-values for both phenotypes are as small or smaller than the observed p-values. Formally, the conjunctional FDR is defined as:

$$ConjFDR(p_1, p_2) = \frac{\pi_0 F_0(p_1, p_2)}{F(p_1, p_2)} + \frac{\pi_1 F_1(p_1, p_2)}{F(p_1, p_2)} + \frac{\pi_2 F_2(p_1, p_2)}{F(p_1, p_2)}$$

where π_0 is the *a priori* fraction of SNPs null for both phenotypes simultaneously, $F_0(p_1, p_2)$ is the joint null cdf, π_1 is the *a priori* fraction of SNPs non-null for the pt1 and null for pt2 with $F_1(p_1, p_2)$ the joint cdf of these SNPs, and π_2 is the *a priori* proportion of SNPs non-null for pt2 and null for pt1, with joint cdf $F_2(p_1, p_2)$. $F(p_1, p_2)$ is the joint overall mixture cdf for all SNPs of both phenotypes. We denote conjunctional FDR for phenotype 1 and phenotype 2 as $FDR_{pt1\&pt2}$.

A model-free conservative estimation of the conjunctional FDR for phenotypes pt1 and pt2 can be calculated as:

$$FDR_{pt1\&pt2} = max\{FDR_{pt1|pt2}, FDR_{pt2|pt1}\}$$

Using upwardly biased estimates of conditional FDRs $FDR_{pt1|pt2}$ and $FDR_{pt2|pt1}$ as described above and noting that for enriched samples, p-values will tend to be smaller than predicted from the uniform distribution (giving $F_1(p_1) \geq p_1$ and $F_2(p_2) \geq p_2$), the latter equation follows from:

$$max\{FDR_{pt1|pt2}, FDR_{pt2|pt1}\} = max\left\{\frac{p_1 F_2(p_2)}{F(p_1, p_2)}, \frac{p_2 F_1(p_1)}{F(p_1, p_2)}\right\}$$

$$\geq (\pi_0 + \pi_1 + \pi_2)max\left\{\frac{p_1 F_2(p_2)}{F(p_1, p_2)}, \frac{p_2 F_1(p_1)}{F(p_1, p_2)}\right\}$$

$$\geq (\pi_0 p_1 p_2 + \pi_1 p_2 F_1(p_1) + \pi_2 p_1 F_2(p_2))/F(p_1, p_2)$$

Assuming that SNPs are independent if one or both are null (that is reasonable for disjoint samples), the last quantity is exactly equal to the formal definition of $ConjFDR(p_1, p_2)$ above.

Identified ADHD associated loci
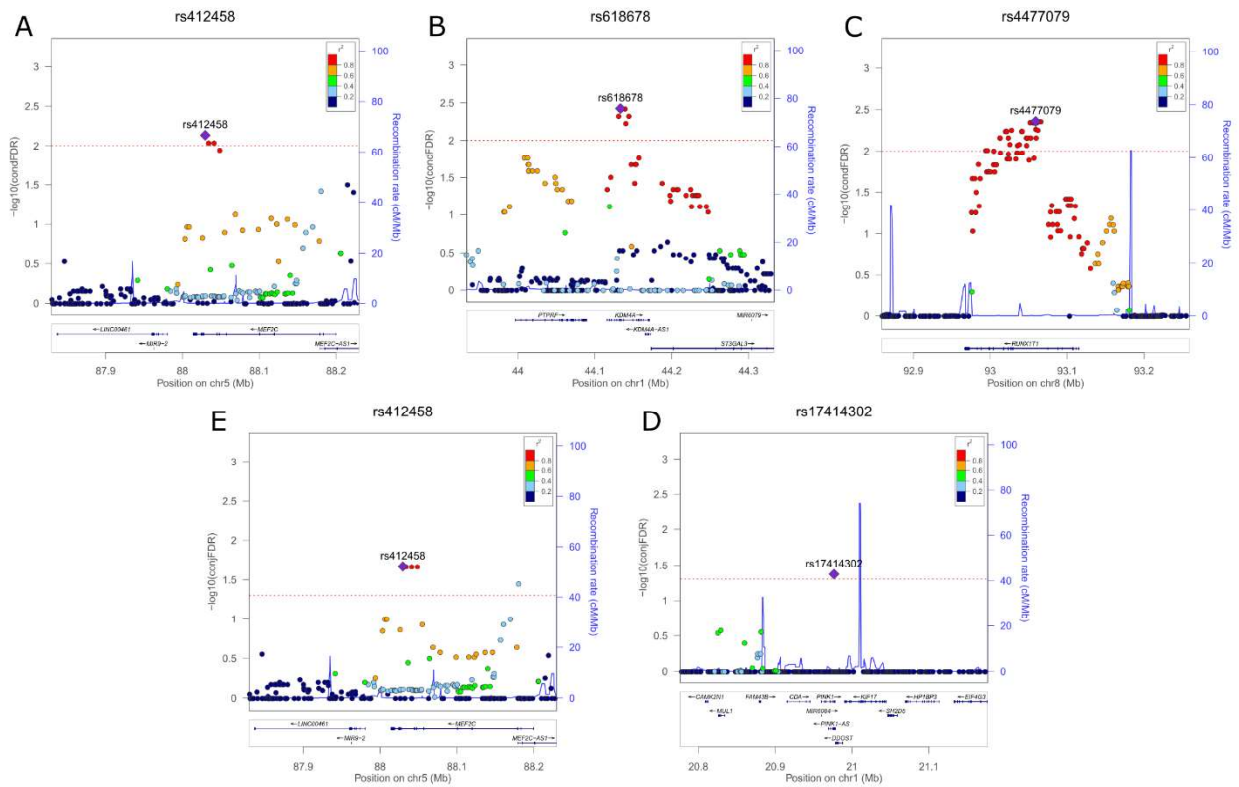
Figure S2. Genetic surrounding of the SNPs identified in the condFDR (upper row, A, B, C) or in the conjFDR (lower row, D, E) analyses except the SNPs with the strongest associations from each analysis (presented in the main text).

Values for both genotyped and imputed variants are shown on the left y-axis as $-\log_{10}$(condFDR) and $-\log_{10}$(conjFDR) respectively. In each subplot, the SNP with the strongest association is shown in the large purple square. The color of the remaining markers reflects the degree of LD with the strongest associated SNP measured as r^2 coefficient (described in the legend). The recombination rate is plotted as a blue solid line, its value in centimorgan/megabase (cM/Mb) is indicated on the right y-axis. Red dotted lines indicate the FDR significance threshold (0.01 for condFDR and 0.05 for conjFDR).

Figures were generated with LocusZoom tool [19] available online: http://locuszoom.sph.umich.edu/locuszoom/

eQTL analyses
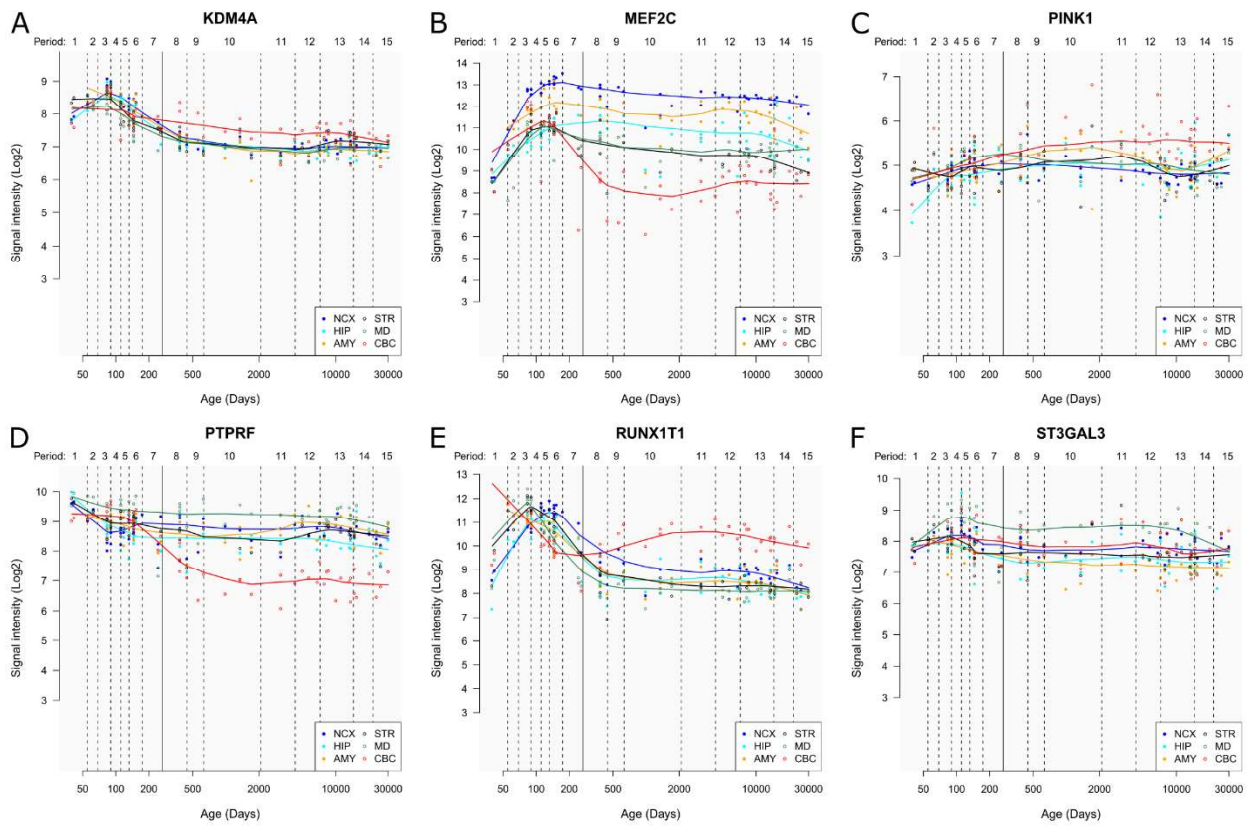
Figure S3. Temporal expression in brain regions for genes from 5 loci identified either in condFDR or in conjFDR analyses. Expression profiles are obtained from the Human Brain Transcriptome database (http://hbatlas.org). Gene expression along entire development and adulthood in the cerebellar cortex (CBC), mediodorsal nucleus of the thalamus (MD), striatum (STR), amygdala (AMY), hippocampus (HIP) and neocortex (NCX) is demonstrated.

Gene	SNP	exprID	aveALL	HIPP	OCTX	PUTM	SNIG	TCTX	THAL	WHMT
ST3GAL3	rs618678	2333515	5.1×10^{-12}	1.8×10^{-5}	7.6×10^{-6}	1.6×10^{-3}	2.1×10^{-3}	4.6×10^{-5}	2.6×10^{-5}	1.6×10^{-2}
PTPRF	rs618678	2333361	9.1×10^{-4}	6.6×10^{-1}	6.7×10^{-1}	1.1×10^{-2}	4.4×10^{-1}	6.6×10^{-2}	2.2×10^{-2}	4.3×10^{-1}
KDM4A	rs618678	2333472	1.3×10^{-2}	3.9×10^{-2}	4.1×10^{-1}	7.9×10^{-1}	1.2×10^{-4}	7.1×10^{-1}	1.2×10^{-1}	3.0×10^{-1}
ERI3	rs618678	2409615	1.6×10^{-2}	3.5×10^{-4}	7.5×10^{-1}	7.4×10^{-1}	8.8×10^{-1}	8.1×10^{-1}	3.2×10^{-2}	3.8×10^{-1}
TIE1	rs618678	2333056	1.0×10^{-1}	6.1×10^{-1}	5.6×10^{-1}	2.1×10^{-4}	4.1×10^{-1}	8.1×10^{-1}	2.4×10^{-1}	1.8×10^{-1}
TIE1	rs618678	2333080	1.8×10^{-1}	5.9×10^{-1}	4.3×10^{-1}	1.7×10^{-1}	8.8×10^{-1}	7.2×10^{-1}	2.8×10^{-4}	4.9×10^{-1}
KIAA0467	rs618678	2333277	3.0×10^{-1}	4.6×10^{-1}	1.4×10^{-1}	5.1×10^{-4}	1.7×10^{-1}	7.8×10^{-1}	7.6×10^{-1}	1
DPH2	rs618678	2333638	5.5×10^{-1}	7.7×10^{-2}	4.4×10^{-2}	2.3×10^{-1}	1.3×10^{-1}	1.9×10^{-1}	5.6×10^{-1}	4.2×10^{-5}
RUNX1T1	rs4477079	3144436	2.6×10^{-1}	9.8×10^{-1}	6.0×10^{-1}	5.9×10^{-4}	9.9×10^{-1}	5.9×10^{-1}	4.1×10^{-1}	6.0×10^{-1}
RUNX1T1	rs4477079	t3144433	2.6×10^{-1}	9.8×10^{-1}	6.0×10^{-1}	5.9×10^{-4}	9.9×10^{-1}	5.9×10^{-1}	4.1×10^{-1}	6.0×10^{-1}
TRIB2	rs4324303	2470209	3.4×10^{-3}	2.3×10^{-1}	2.2×10^{-1}	1.5×10^{-1}	5.2×10^{-1}	1.2×10^{-4}	5.8×10^{-1}	3.8×10^{-2}
TMCO4	rs17414302	2399954	5.2×10^{-4}	1.3×10^{-1}	5.7×10^{-1}	9.6×10^{-1}	2.8×10^{-2}	1.2×10^{-1}	2.9×10^{-1}	4.3×10^{-1}
TMCO4	rs17414302	2399978	9.6×10^{-4}	1.3×10^{-2}	6.0×10^{-1}	5.5×10^{-1}	6.0×10^{-2}	5.8×10^{-1}	2.5×10^{-2}	3.1×10^{-1}
TMCO4	rs17414302	t2399977	9.6×10^{-4}	1.3×10^{-2}	6.0×10^{-1}	5.5×10^{-1}	6.0×10^{-2}	5.8×10^{-1}	2.5×10^{-2}	3.1×10^{-1}
PINK1	rs17414302	2324112	6.4×10^{-2}	5.8×10^{-3}	4.6×10^{-1}	5.4×10^{-2}	8.5×10^{-1}	6.0×10^{-1}	4.6×10^{-1}	5.5×10^{-4}
TMCO4	rs17414302	2399969	9.8×10^{-2}	7.4×10^{-1}	6.0×10^{-1}	2.1×10^{-1}	5.6×10^{-4}	7.6×10^{-1}	6.3×10^{-1}	7.7×10^{-1}
RAP1GAP	rs17414302	2400695	1.9×10^{-1}	5.2×10^{-4}	4.3×10^{-1}	8.8×10^{-1}	5.1×10^{-1}	1.4×10^{-1}	2.0×10^{-1}	1.2×10^{-1}
EIF4G3	rs17414302	2400449	3.4×10^{-1}	4.6×10^{-1}	5.1×10^{-1}	7.6×10^{-4}	1.2×10^{-1}	7.6×10^{-1}	7.0×10^{-1}	2.3×10^{-1}
RAP1GAP	rs17414302	2400691	7.6×10^{-1}	8.9×10^{-4}	4.8×10^{-1}	8.1×10^{-1}	5.6×10^{-1}	5.6×10^{-1}	2.8×10^{-1}	4.2×10^{-1}

Table S1. Genes with expression levels significantly associated with one of SNPs identified as significant either in condFDR or in conjFDR analyses (Table 1) according to Braineac database (http://www.braineac.org). First column (Gene) represents the gene symbol. Second column (SNP) shows the rs number of the SNP that affects gene expression. Third column (exprID) indicates the identification number of exon-specific probeset affected by the SNP (if ID starts with "t", expression is affected on the transcript level otherwise effect is detected on the exon level). Fourth column (aveALL) demonstrates p-value for average expression across all 10 brain tissues available in Braineac. The remaining 7 columns contain eQTL p-values for 7 different brain regions: HIPP (hippocampus), occipital OCTX (cortex, specifically primary visual cortex), PUTM (putamen), SNIG (substantia nigra), TCTX (temporal cortex), THAL (thalamus) and WHMT (intralobular white matter). Significant p-values (< 0.001) are highlighted in red. The remaining 3 brain regions available in Braineac: cerebellar cortex, frontal cortex and medulla (specifically inferior olivary nucleus) are not included in the table because none of the genes affected by identified SNPs have significantly altered expression level in these tissues.

Evaluation of the detected ADHD loci in an independent study of ADHD symptoms

SNP	Chr region	Position	Effect size		p-value	
			EAGLE	PGC	EAGLE	PGC
rs17414302	1p36.12	20976535	-0.293	-0.090	7.69×10^{-1}	1.97×10^{-4}
rs618678	1p34.2	44133299	-2.364	-0.053	1.80×10^{-2}	1.05×10^{-5}
rs4324303	2p24	13817678	-0.827	-0.079	4.08×10^{-1}	2.05×10^{-7}
rs412458	5q14.3	88029627	0.230	0.061	8.18×10^{-1}	2.15×10^{-5}
rs4477079	8q21.3	93059038	0.170	-0.071	8.65×10^{-1}	1.44×10^{-6}

Table S2. The most significant SNPs in the loci identified by condFDR/conjFDR, their effect sizes and association p-values in PGC ADHD GWAS and EAGLE GWAS of ADHD symptoms.

Supplemental references

1. Neale BM, Medland SE, Ripke S, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry.* 2010;49(9):884-897.
2. Elia J, Gai X, Xie HM, et al. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry.* 2010;15(6):637-646.
3. Neale BM, Lasky-Su J, Anney R, et al. Genome-wide association scan of attention deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet.* 2008;147B(8):1337-1344.
4. Neale BM, Medland S, Ripke S, et al. Case-control genome-wide association study of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry.* 2010;49(9):906-920.
5. Mick E, Todorov A, Smalley S, et al. Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry.* 2010;49(9):898-905 e893.
6. Andreassen OA, Djurovic S, Thompson WK, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet.* 2013;92(2):197-209.
7. Andreassen OA, Thompson WK, Schork AJ, et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet.* 2013;9(4):e1003455.
8. Okbay A, Beauchamp JP, Fontana MA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature.* 2016;533(7604):539-542.
9. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010;26(17):2190-2191.
10. Middeldorp CM, Hammerschlag AR, Ouwens KG, et al. A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry.* 2016;55(10):896-905 e896.

11. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th Ed).* Washington DC: Washington DC : American Psychiatric Association, 2013; 2013.

12. Achenbach TMR, L.A. *Manual for the ASEBA Preschool forms and Profiles.* University of Vermont, Research Center for Children, Youth, and Families: Burlington, VT.; 2000.

13. Achenbach TMR, L.A. *Manual for the ASEBA School-Age Forms and Profiles.* University of Vermont, Research Center for Children, Youth, and Families: Burlington, VT.; 2001.

14. Goodman R. The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry.* 1997;38(5):581-586.

15. Conners CK. *Conner's rating scales-revised: Technical manual.* Toronto: Multi-Health Systems; 1997.

16. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-1241.

17. Andreassen OA, Thompson WK, Schork AJ, et al. Correction: Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLoS Genet.* 2015;11(11):e1005544.

18. Efron B. *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction.* Cambridge ; New York: Cambridge University Press; 2010.

19. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336-2337.