Identification of Genetic Loci Shared Between Attention-Deficit/Hyperactivity Disorder, Intelligence and Educational Attainment

Supplement 1

Supplementary Methods

GWAS Samples

GWAS summary statistics for ADHD were obtained from the Psychiatric Genomics Consortium (PGC) and comprised association analyses of 19,099 cases and 34,194 controls (n total = 53,293) (1). These samples included a population-based cohort of 14,584 cases and 22,492 controls from Denmark collected by the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), and 10 European cohorts aggregated by the Psychiatric Genomics Consortium (PGC) (1). ADHD cases in iPSYCH were identified from a national research register and diagnosed by psychiatrists at a psychiatric hospital according to ICD10 (F90.0), and genotyped using Illumina PsychChip. The PGC cohorts include both adult and child participants. Further detailed descriptions of all included cohorts, including sex distributions, have been described previously (2).

The summary statistics for general intelligence (INT) were obtained from the meta-analysis of 14 independent cohorts (n = 269,867), comprised of adult and child participants of European ancestry (3). For general intelligence, the INT GWAS included cohorts assessed using various measures of intelligence (3). Statistically, the variance common across cognitive tasks can be labeled general intelligence or Spearman's *g* (4). The g factors extracted from different sets of cognitive tests are known to correlate very strongly (5), and were used as the outcome variable in this GWAS (3). Furthermore, a detailed description of these cohorts is available in the original publication (3).

For our analyses of educational attainment (EDU) we used summary statistics (n = 842,499) generated from the meta-analysis of data from the Social Science Genetic Association Consortium (SSGAC) (n = 766,344) (6) and 23andMe (n = 76,155) (7). Data for the additional 23andMe participants (n = 289,383) included in the Lee et al. study (6) were not available for the present study. The meta-analysis was performed using an inverse-weighted fixed effects model implemented in the software METAL (http://csg.sph.umich.edu//abecasis/Metal/) (8). Each major educational qualification outcome from the surveys used in the cohorts included in the EDU meta-analysis were mapped to an International Standard Classification of Education (ISCED) category (6, 7). These category scores were then imputed to obtain a 'years of education' score (1 – 22 years) that was then used as the outcome variable in the meta-analysis (6). All participants included in the meta-analysis were adults of European ancestry, and additional detailed descriptions of the included participants are provided in the original publication (6).

Independent Study of ADHD Symptoms – EAGLE Cohort

The Early Genetics and Lifecourse Epidemiology (EAGLE) consortium includes population-based cohorts from Europe, Australia and the United States. For this study of ADHD symptoms, nine EAGLE cohorts were included with available scores in childhood (age at measurement < 13 years). Further details of the nine cohorts is provided in the original article (9). In order to assess ADHD symptoms, different instruments were used across cohorts, including the Attention Problems scale of the Child Behavior Checklist (CBCL) and the Teacher Report Form (TRF), and the Hyperactivity scale of the Strengths and Difficulties Questionnaire (SDQ). For the meta-analysis, one phenotype was selected from each cohort. A detailed description of the quality control, imputation and the analysis procedures for the different cohorts can be found in

the original article (9). Association analyses were performed using linear regression and relevant principal components and subsequently meta-analysed using METAL (8). Summary statistics from the meta-analysis of N=17,666 individuals were used in the current study.

Independent ADHD Case-Control Sample – deCODE Genetics, ADHD Diagnoses from Medical Records

The Icelandic ADHD cohort (n = 10,217) is comprised of individuals who have either a clinical ADHD diagnosis (mostly ICD10-F90) or who have been prescribed medication specific for ADHD symptoms (ATC-NA06BA, mostly methylphenidate). The Icelandic control sample (n = 338,344) consists of individuals participating in various deCODE studies. All individuals used in the analysis were subject to chip genotyping and long range phasing and genotypes were imputed based on the Icelandic dataset as described previously (10).

Conditional False Discovery Rate

The 'enrichment' seen in the conditional QQ plots and fold-enrichment plots can be directly interpreted in terms of true discovery rate (TDR = 1 – false discovery rate (FDR)) (11). More specifically, for a given p-value cutoff, the FDR is defined as

$$FDR(p) = \pi_0 F_0(p) / F(p), \qquad [1]$$

where π_0 is the proportion 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 (12). Under the null hypothesis, F_0 is the cdf of the uniform distribution on the unit interval [0,1], so that Eq. [1] reduces to

$$FDR(p) = \pi_0 \, p / F(p), \qquad [2]$$

The cdf F can be estimated by the empirical cdf $q = N_p / N$, where N_p is the number of SNPs with p-values less than or equal to p, and N is the total number of SNPs. Replacing F by q in Eq. [2], we get

$$\text{Estimated FDR}(p) = \pi_0 \, p / q, \qquad [3]$$

which is biased upwards as an estimate of the FDR (13). Replacing π_0 in Equation [3] with unity gives an estimated FDR that is further biased upward;

$$q^* = p/q, \qquad [4]$$

If π_0 is close to one, as is likely true for most GWASs, the increase in bias from Eq. [3] is minimal. The quantity 1 – p/q, is therefore biased downward, and hence a conservative estimate of the TDR. Referring to the QQ plots, we see that q* is equivalent to the nominal p-value divided by the empirical quantile, as defined earlier. We can thus read the FDR estimate directly off the QQ plot as

$$-\log_{10}(q^*) = \log_{10}(q) - \log_{10}(p), \qquad [5]$$

i.e. the horizontal shift of the curves in the QQ plots from the expected line x = y, with a larger shift corresponding to a smaller FDR (Illustrated in Figure 1 and Supplementary Figure S1).

Conditional QQ Plots

Under large-scale testing paradigms, such as GWAS, quantitative estimates of likely true associations can be drawn from the distributions of summary statistics (12, 14). One common method for visualizing the enrichment of statistical association relative to that expected under the global null hypothesis is through QQ plots of the nominal p-values obtained from GWAS summary statistics. QQ plots compare a nominal probability distribution against an empirical distribution. In the presence of all null relationships, nominal p-values form a straight line on a QQ plot when plotted against the empirical distribution. Under the global null hypothesis the theoretical p-value distribution is uniform on the interval [0,1]. As is common in GWAS, we plot $-\log_{10}$ p against $-\log_{10}$ q (q=1-cdf(p)) to emphasize tail probabilities of the theoretical and empirical distribution of ADHD, EDU and INT associations. Leftward deflections of the observed distribution from the projected null line reflect increased tail probabilities in the

distribution of test statistics (z-scores) and consequently an over-abundance of low p-values compared to that expected by chance, also named 'enrichment'.

Conditional QQ plots are constructed by creating subsets of SNPs based on levels of an auxiliary measure for each SNP, and computing QQ plots separately for each level. If SNP enrichment is captured by variation in the auxiliary measure, this is expressed as successive leftward deflections in a conditional QQ plot as levels of the auxiliary measure increase. We constructed conditional QQ plots of empirical quantiles of nominal association $-\log_{10}$ p-values for all SNPs and for subsets (strata) of SNPs determined by the nominal p-values of their association with the conditional phenotypes, and vice versa. Specifically, we computed the empirical cumulative distribution (cdf) of nominal p-values for a given phenotype for all SNPs and for SNPs with significance levels below the indicated cut-offs for the conditional phenotypes ($-\log_{10}(p) \geq 1$, $-\log_{10}(p) \geq 2$, $-\log_{10}(p) \geq 3$ corresponding to $p < 0{\cdot}1$, $p < 0{\cdot}01$, $p < 0{\cdot}001$ respectively). To assess the polygenic effects below the standard GWAS significance threshold, we focused the conditional QQ plots on SNPs with nominal $-\log_{10}(p) < 7{\cdot}3$ (corresponding to $p > 5x10^{-8}$). To control for spurious enrichment, all conditional QQ plots were constructed after random pruning averaged over 500 iterations. At each iteration, one SNP in every LD block (defined by an $r^2 > 0{\cdot}1$) was randomly selected and the empirical cdfs were computed using the corresponding p-values.

Detection of Genetic Variants Using Conditional and Conjunctional FDR

The FDR can be interpreted as the probability that a SNP is null given that its p-value is as small as or smaller than its observed p-value. The conditional FDR (condFDR) is an extension of the standard FDR, which incorporates information from GWAS summary statistics of a second phenotype to adjust its significance level. The condFDR is defined as the probability

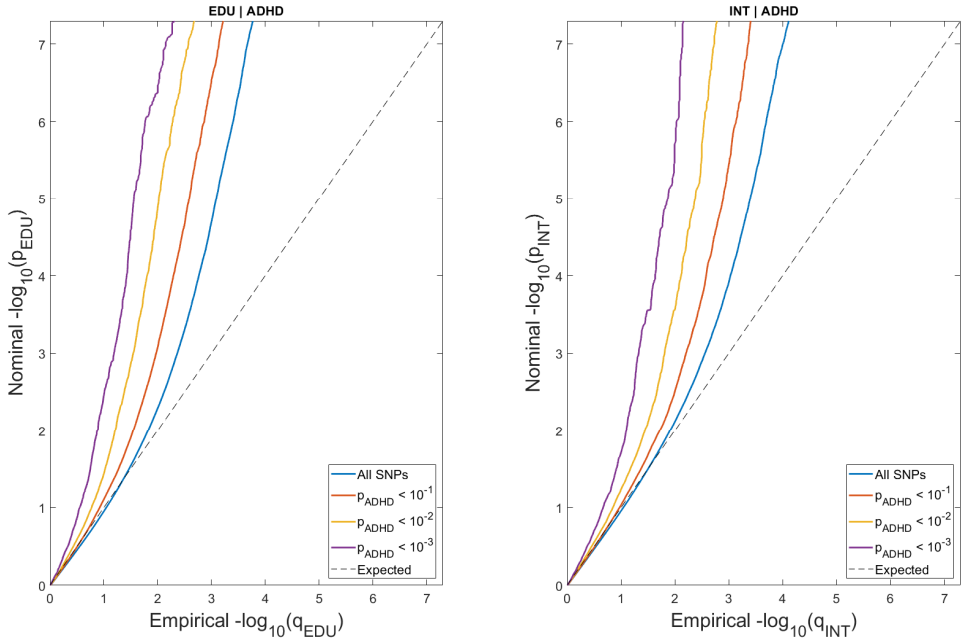
that a SNP is null in the first phenotype given that the p-values in the first and second phenotypes are as small as or smaller than the observed ones. Ranking SNPs by the standard FDR or by p-values gives the same ordering of SNPs. In contrast, if the primary and secondary phenotypes are related genetically, the condFDR reorders SNPs and results in a different ranking than that based on p-values alone. The conjunctional FDR (conjFDR) is defined as the posterior probability that a SNP is null for either phenotype or both simultaneously, given that its p-values for association with both phenotypes are as small as or smaller than the observed p-values (15, 16). A conservative estimate of the conjFDR is given by the maximum condFDR for a given SNP after repeating the condFDR procedure for both traits and inverting their roles (17). Given that complex correlations in regions with intricate LD can bias FDR estimation (18), we excluded SNPs in the extended major histocompatibility complex (genome build 19 locations 25119106–33854733) and SNPs in LD (r^2>0·1) with such SNPs before fitting the FDR models.

Code Availability

All of the code necessary to replicate the condFDR/conjFDR analyses described above is publicly available at https://github.com/precimed/pleiofdr.

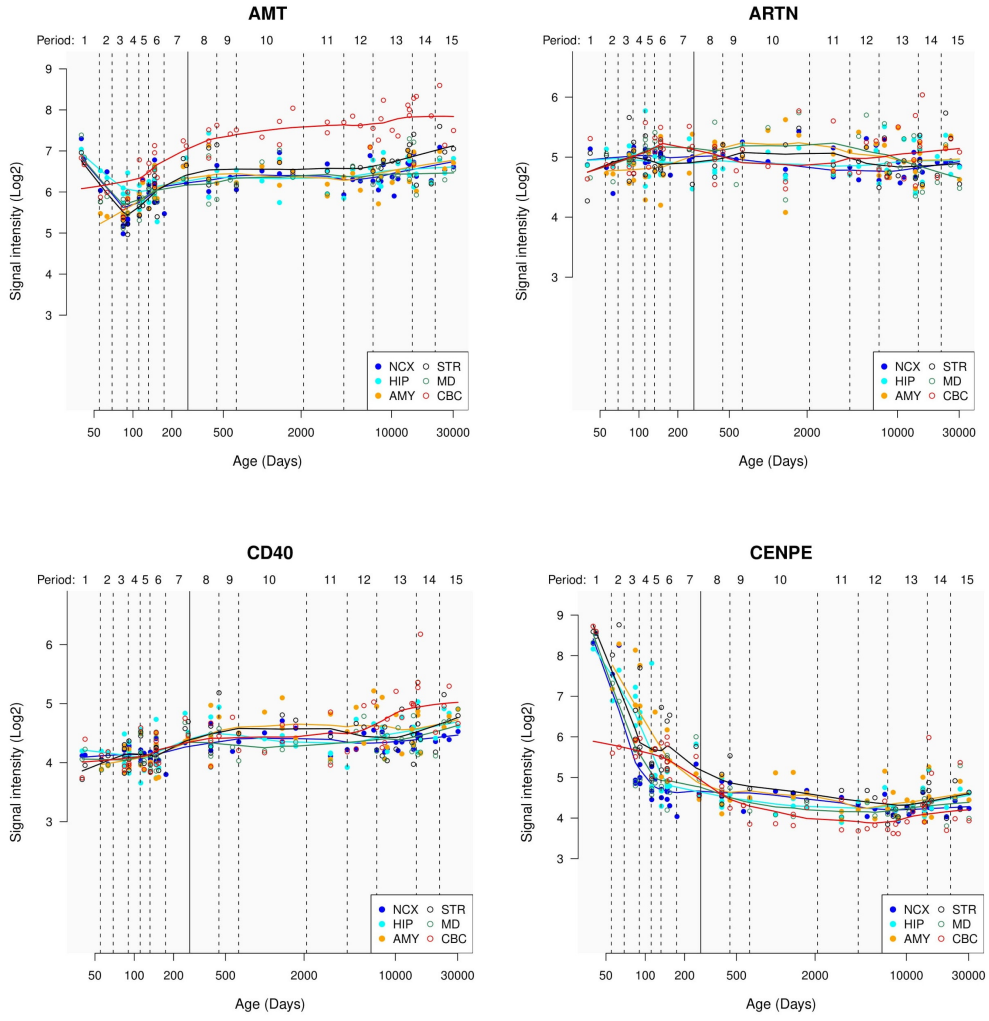
Supplementary Figures

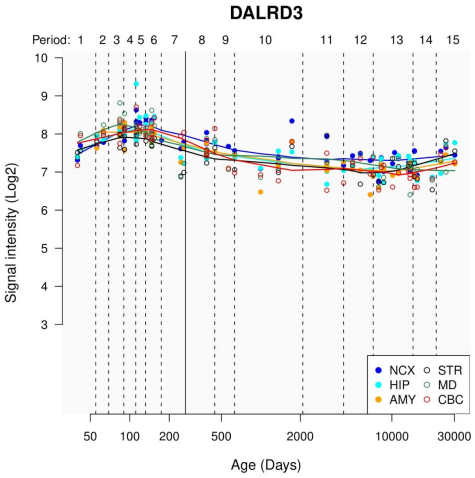
Supplementary Figure S1. Stratified conditional QQ plots of nominal vs empirical −log10 p-values (corrected for inflation) in educational attainment (EDU) or general intelligence (INT) below the standard genome-wide association study threshold of $p < 5 \times 10^{-8}$ as a function of significance of association with ADHD at the level of −log10 p-values of 1, 2, or 3, corresponding to p = 0.10, p = 0.01 and p = 0.001, respectively. The dashed lines indicate the null hypothesis.

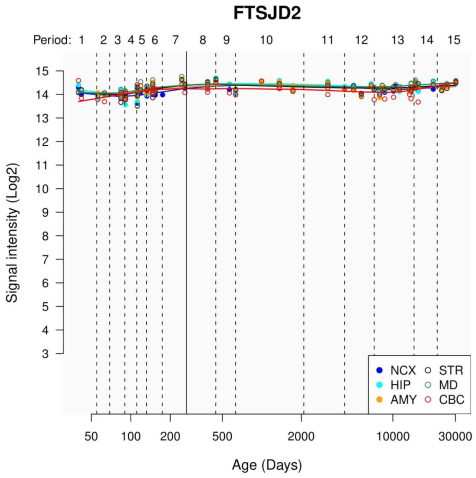


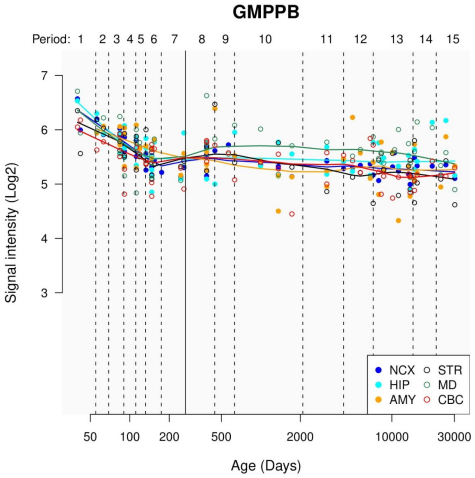
Supplementary Figure S2. Age-dependent variations of expression for the genes associated with identified eQTL SNPs in the developmental and adult human brain.

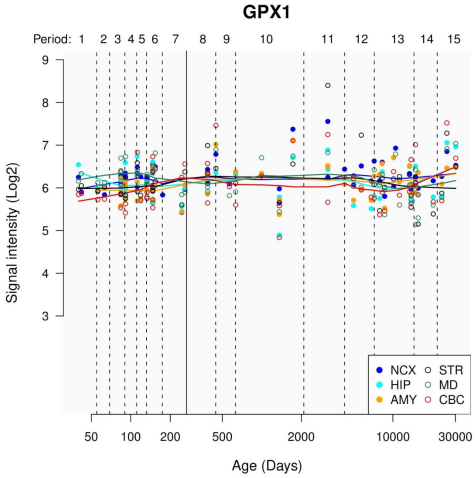
Line plots show the log2-transformed gene exon array signal intensity from the early fetal period to late adulthood in six brain regions. The solid line between periods 7 and 8 (approximately post-conception day 280) separates prenatal from postnatal periods. Data were generated using Affymetrix GeneChip Human Exon 1.0 ST Arrays by the Human Brain Transcriptome project, and accessed via their publicly available database at http://hbatlas.org. Abbreviations: NCX = neocortex; HIP = hippocampus; AMY = amygdala; STR = striatum; MD = mediodorsal nucleus of the thalamus; CBC = cerebellar cortex.

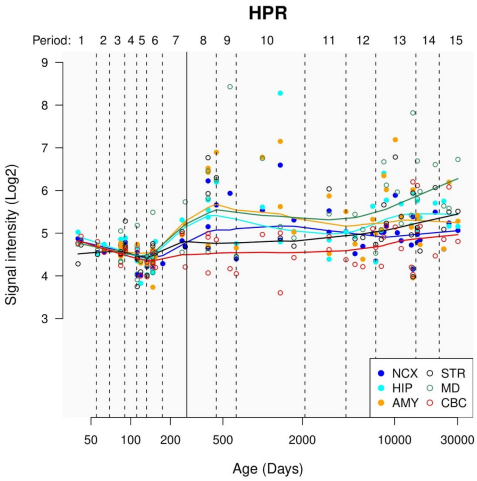


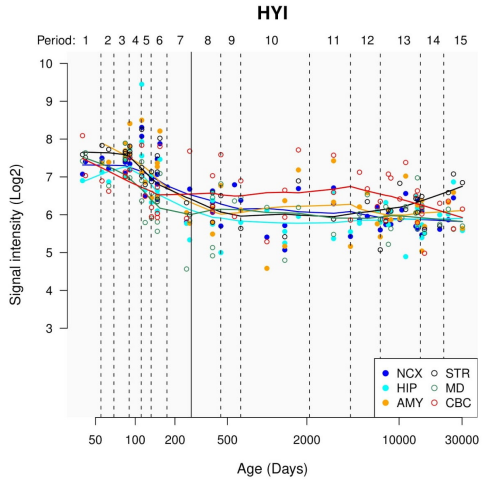

DALRD3
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

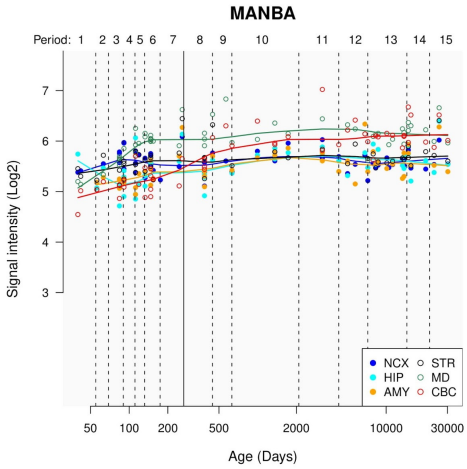

FTSJD2
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

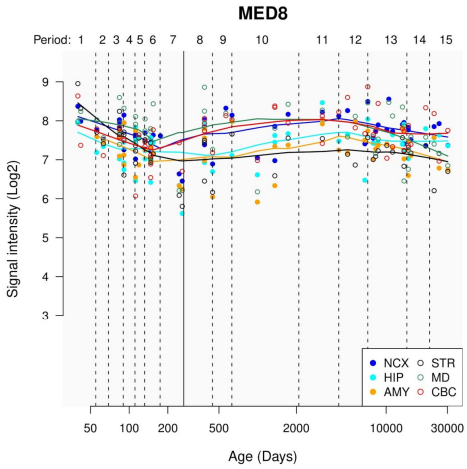

GMPPB
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

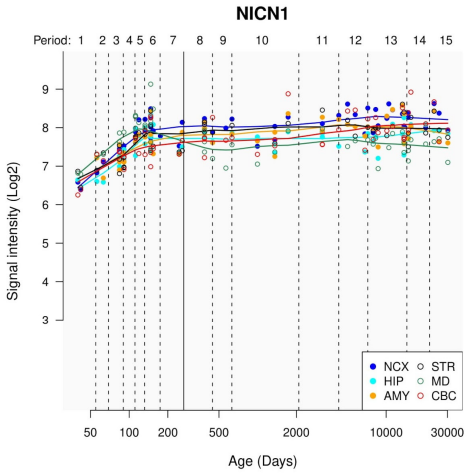

GPX1
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

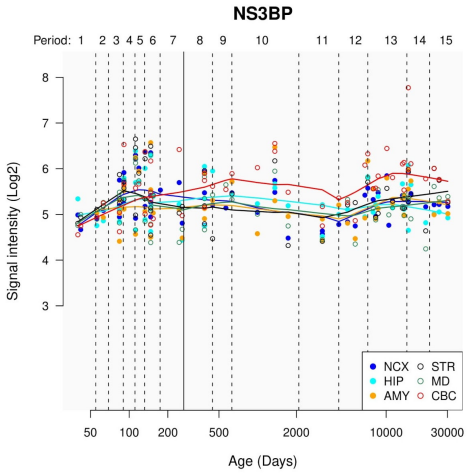

HPR
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

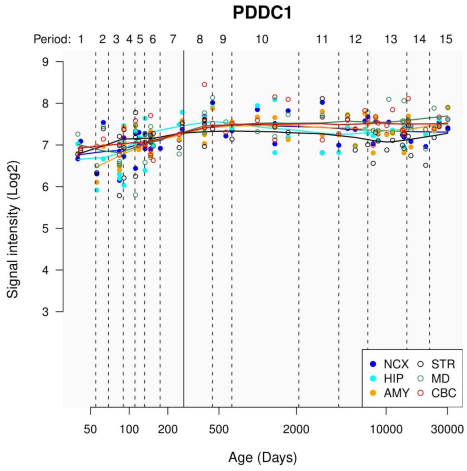

HYI
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
50 100 200 500 2000 10000 30000
NCX STR
HIP MD
AMY CBC

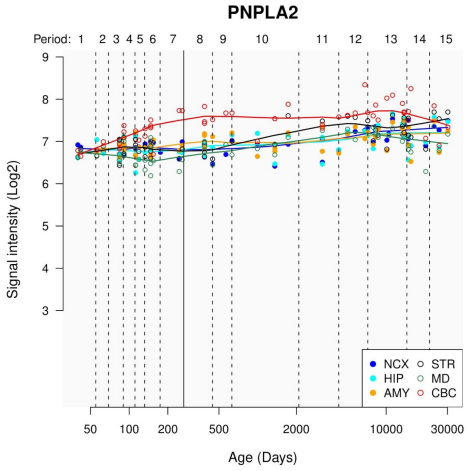

MANBA
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)

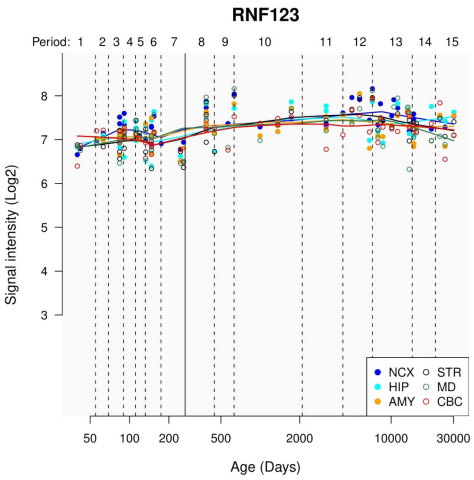

MED8
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7 8 9
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)

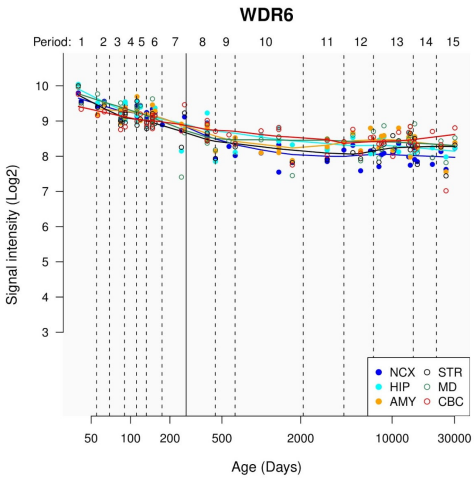

NICN1
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7 8 9
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)

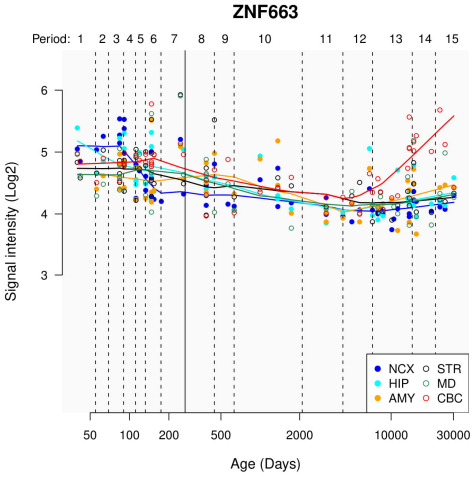

NS3BP
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7 8
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)


PDDC1
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7 8 9
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)


PNPLA2
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
3 4 5 6 7 8 9
NCX STR
HIP MD
AMY CBC
50 100 200 500 2000 10000 30000
Age (Days)


RNF123
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
NCX STR
HIP MD
AMY CBC


WDR6
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
NCX STR
HIP MD
AMY CBC


ZNF663
Period: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Signal intensity (Log2)
Age (Days)
NCX STR
HIP MD
AMY CBC

Supplemental References

1. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, *et al.* (2019): Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet*. 51: 63–75.

2. Neale BM, Medland SE, Ripke S, Asherson P, Franke B, Lesch K-P, *et al.* (2010): Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*. 49: 884–897.

3. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, *et al.* (2018): Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet*. 50: 912–919.

4. Spearman C (1904): "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*. 15: 201–292.

5. Johnson W, Nijenhuis J te, Bouchard TJ (2008): Still just 1 g: Consistent results from five test batteries. *Intelligence*. 36: 81–95.

6. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, *et al.* (2018): Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet*. 50: 1112–1121.

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

8. Willer CJ, Li Y, Abecasis GR (2010): METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 26: 2190–2191.

9. Middeldorp CM, Hammerschlag AR, Ouwens KG, Groen-Blokhuis MM, Pourcain BS, Greven CU, *et al.* (2016): A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts. *J Am Acad Child Adolesc Psychiatry*. 55: 896-905.e6.

10. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, *et al.* (2015): Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*. 47: 435–444.

11. Benjamini Y, Hochberg Y (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 57: 289–300.

12. Efron B (2007): Size, power and false discovery rates. *Ann Statist*. 35: 1351–1377.

13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007): PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 81: 559–575.

14. Schweder T, Spjøtvoll E (1982): Plots of P-values to evaluate many tests simultaneously. *Biometrika*. 69: 493–502.

15. Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, *et al.* (2013): Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet*. 92: 197–209.

16. Andreassen OA, Thompson WK, Dale AM (2013): Boosting the Power of Schizophrenia Genetics by Leveraging New Statistical Tools. *Schizophr Bull*. 40: 13–17.

17. Nichols T, Brett M, Andersson J, Wager T, Poline J-B (2005): Valid conjunction inference with the minimum statistic. *Neuroimage*. 25: 653–660.

18. Schwartzman A, Lin X (2011): The effect of correlation in false discovery rate estimation. *Biometrika*. 98: 199–214.