Shared Genetics of Asthma and Mental Health Disorders: A Large-Scale Genome-Wide Cross-Trait Analysis

Zhaozhong Zhu, Sc.D., Xi Zhu, M.D., Cong-Lin Liu, M.D., Ph.D., Huwenbo Shi, Ph.D., Sipeng Shen, Yunqi Yang, Kohei Hasegawa, M.D., M.P.H., Carlos A. Camargo, Jr., M.D., Dr.P.H., Liming Liang, Ph.D.

Online supplementary note

UK Biobank dataset

The UK Biobank study is a prospective study of >500,000 participants living in the UK. In total, 503,325 participants who registered in the National Health Service with ages ranging 40–69 years were recruited out of 9.2 million mailed invitations. Baseline data were collected using questionnaires, and anthropometric assessments were performed. All detailed genotyping, quality control, and imputation procedures are described at the UK Biobank website (http://biobank.ctsu.ox.ac.uk). Currently ~500,000 individuals in UK Biobank have been genotyped for ~800,000 SNPs using Affymetrix facilities. Population structure was captured by principal component analysis on ~500,000 UK Biobank samples using ~100,000 SNPs. Quality control filters were applied before phasing. Data were prephased using SHAPEIT3 [1]. Haplotype Reference Consortium (HRC) panel data was used as a reference panel for imputation. This reference panel has many more haplotypes (64,976) than the 1000G reference panel, and so is expected to produce better imputation performance [2].

We used 4 data fields to determine asthmatic cases: 6152, 20002, 41202 and 41204. Data field 6152 is from the participant questionnaire to determine the doctor-diagnosed asthma phenotypes. This data field contains the question: “Has a doctor ever told you that you have had any of the following conditions?” Participants could select more than one answer from the following: Blood clot in the leg (DVT); blood clot in the lung; emphysema/chronic bronchitis; asthma; hayfever, allergic rhinitis or eczema; none of the above; prefer not to answer. If participants chose either "none of the above" or "prefer not to answer", they could not select other answers. Data field 20002 denotes self-reported non-cancer illness code. Data field 41202 and 41204 denote ICD10 main and secondary diagnoses from hospital.

We used data fields 3786 (age of first asthma was diagnosed based on touchscreen questionaire) and 22147 (age of first asthma was diagnosed by doctor based on an online follow-up questionnaire finished only by subset of participants) to determine the asthma age of onset. Three asthma subtypes were used in this study: childhood-onset asthma (defined as asthma age of onset [AAO]≤12 years old), adult-onset asthma (AAO≥26) and young adult-onset asthma (12<AAO<25). The young adult-onset asthma was not included in the genetic analysis due to its higher heterogeneity. Since we used both data fields 3786 and 22147 to determine AAO, we performed addition quality control to handle inconsistencies between these 2 data fields. Specifically, we first excluded 8,307 subjects with missing AAO for both data fields; then we excluded 426 subjects with data fields 3786 and 22147 for AAO inconsistency > 10 years; finally we excluded 441 subjects with AAO inconsistency ≤ 10 years but have inconsistent age group from these 2 data fields.

To assess phenotypic correlation between asthma and mental health disorders in UK Biobank, we additionally extracted phenotypes from UK Biobank, including depression (MDD) (data fields 20002, 20126, 20544), anxiety (ANX) (20002, 20544, 20544, 41202, 41204), posttraumatic stress disorder (PTSD) (20002, 41202, 41204), bipolar disorder (BIP) (20002, 20126, 20544, 41202, 41204), eating disorder (ED) (20002, 20544, 41202, 41204) and schizophrenia SCZ) (20002, 20544, 41202, 41204).

All participants from this study provided UK Biobank-acquired informed consent and provided data according to the UK Biobank protocol. We have complied with all ethical regulations according to UK Biobank policy. This research was approved and conducted using the UK Biobank under application number 16549 and 45052.

Attention Deficit Hyperactivity Disorder (ADHD) dataset

An international collaborative team including the Psychiatric Genomics Consortium (PGC) conducted a meta-analysis of GWAS of individuals with ADHD. Participants included both children and adults. A European subset of GWAS data was used in current study. Key summary information can be found in Table S1.

ANX dataset

The Anxiety NeuroGenetics STudy (ANGST) Consortium conducted a meta-analysis of GWAS of individuals with ANX and controls. All participants were adults and European ancestry. Key summary information can be found in Table S1.

ASD dataset

The ASD working group of the PGC conducted a meta-analysis of GWAS of individuals with ASD and controls. Participants included both children and adults. A European subset of GWAS data was used in current study. Key summary information can be found in Table S1.

BIP dataset

The BIP working group of the PGC conducted a meta-analysis of GWAS of individuals with bipolar disorder and controls. Participants included both children and adults. All subjects from this study are European ancestry. Key summary information can be found in Table S1.

ED dataset

This study is based on a meta-analysis of GWAS of individuals with ED and controls. Participants included both children and adults. All subjects from this study are European ancestry. Key summary information can be found in Table S1.

MDD dataset

The MDD working group of the PGC conducted a meta-analysis of GWAS of individuals with MDD and controls. Participants included both children and adults, but majority are adults. A European subset of GWAS data excluding 23andme was used in current study. However, for Mendelian randomization analysis, 10K top significant SNPs including 23andme samples were used. Key summary information can be found in Table S1.

PTSD dataset

The PTSD working group of PGC conducted a meta-analysis of GWAS of individuals with PTSD and controls. All participants were adults. A European subset of GWAS data was used in current study. Key summary information can be found in Table S1.

SCZ dataset

The SCZ working group of the PGC conducted a meta-analysis of GWAS of individuals with SCZ and controls. Although the meta-analysis of 49 cohorts contains 2 ancestries, majority of them are from European ancestry (46 of European and three of east Asian ancestry, 34,241 cases and 45,604 controls). These comprise the primary PGC GWAS data set. Participants included both children and adults. Key summary information can be found in Table S1.

GWAS analysis of UK Biobank data

To account for relatedness, the association between cardiac traits in UK Biobank data and imputed SNPs was carried out using BOLT-linear mixed model (LMM) [3]. The output of BOLT-LMM linear regression was transformed into log odds ratio (logOR) for HBP binary phenotype using the following equation:

$$logOR = \frac{Beta_{BOLT-LMM}}{\frac{N_{case}}{N_{control}} * \left(1 - \frac{N_{case}}{N_{control}}\right)}$$

Association analysis based on subsets (ASSET)

ASSET is a generalized fixed-effects meta-analysis model that combines effect estimate and standard error of GWAS of related but distinct traits to identify promising directions to discover loci with small but common pleiotropic effects. The ASSET method explores subsets of studies for the presence of true association signals that are either in the same direction or opposite directions [4]. When *S* represents a set of study traits selected from *K* studies, meta-analysis statistics of the one-sided test ASSET is defined as:

$$\mathrm{Z}_{\mathrm{max-ASSET}} = max_{S \in \boldsymbol{S}} \mid Z(S)| \ = max_{S \in \boldsymbol{S}} | \sum_{k \in S} \sqrt{\pi_k(S)} Z_k \mid$$

where *S* is all possible 2^K–1 subsets of K studies, and $\pi_k(T) = n_k / \sum_{k \in T} n_k$ represents sample size of the study *K* relative to total sample size of the given subset *S*.

An advantage of using ASSET is that it can account for correlation among studies/subjects that might arise due to shared subjects across distinct studies or due to correlation among related traits in the same study by using case–control overlap matrices. If *Z*(*A*) and *Z*(*B*) denote *Z* statistics for the association test for a SNP from case–control studies A and B with an arbitrary amount of overlap between subjects, then—under the null hypothesis of no association and the assumption that there is no covariate adjustment—the correlation between statistics is given by

$$Corr\{Z(A), Z(B)\} = \sqrt{\frac{n_A^{(1)} n_A^{(0)}}{N_A}} \sqrt{\frac{n_B^{(1)} n_B^{(0)}}{N_B}} \left[\frac{n_{AB}^{(11)}}{n_A^{(1)} n_B^{(1)}} - \frac{n_{AB}^{(10)}}{n_A^{(1)} n_B^{(0)}} - \frac{n_{AB}^{(01)}}{n_A^{(0)} n_B^{(1)}} + \frac{n_{AB}^{(00)}}{n_A^{(0)} n_B^{(0)}} \right]$$

where $n_A^{(1)}$, $n_A^{(0)}$, and N_A are the number of cases, controls, and subjects, respectively, in study A; $n_B^{(1)}$, $n_B^{(0)}$, and N_B are the number of cases, controls, and subjects, respectively, in study B; and $n_{AB}^{(ij)}$ represents the number of subjects with different phenotype categories (i,j) ∈ (0,1) that overlap between studies A and B. For example, $n_{AB}^{(11)}$ denotes the number of shared cases between studies A and B; $n_{AB}^{(10)}$ denotes the number of individuals who are treated as cases in study A but as controls in study B; $n_{AB}^{(01)}$ denotes the number of individuals who are treated as controls in study A but as cases in study B; and $n_{AB}^{(00)}$ denotes the number of shared controls between studies A and B [4].

Reference

1. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, Zagury JF, Delaneau O, Marchini J. Haplotype estimation for biobank-scale data sets. *Nat Genet* 2016: 48(7): 817-820.

2. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PI, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R, Haplotype Reference C. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016: 48(10): 1279-1283.

3. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger BJNg. Efficient Bayesian mixed-model analysis increases association power in large cohorts. 2015: 47(3): 284.

4. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee NJTAJoHG. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. 2012: 90(5): 821-835.