ARTICLES
https://doi.org/10.1038/s41588-020-00767-x

nature genetics

Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program

Murray B. Stein[1,2,3,15]✉, Daniel F. Levey[4,5], Zhongshan Cheng[4,5], Frank R. Wendt[4,5], Kelly Harrington[6,7], Gita A. Pathak[4,5], Kelly Cho[6,8], Rachel Quaden[6], Krishnan Radhakrishnan[9,10,11], Matthew J. Girgenti[4,5], Yuk-Lam Anne Ho[6], Daniel Posner[6], Mihaela Aslan[9,12], Ronald S. Duman[4,5,16], Hongyu Zhao[9,13], Department of Veterans Affairs Cooperative Studies Program (no. 575B)*,**, VA Million Veteran Program*,**, Renato Polimanti[4,5], John Concato[9,12,14] and Joel Gelernter[4,5,15]✉

We conducted genome-wide association analyses of over 250,000 participants of European (EUR) and African (AFR) ancestry from the Million Veteran Program using electronic health record-validated post-traumatic stress disorder (PTSD) diagnosis and quantitative symptom phenotypes. Applying genome-wide multiple testing correction, we identified three significant loci in European case-control analyses and 15 loci in quantitative symptom analyses. Genomic structural equation modeling indicated tight coherence of a PTSD symptom factor that shares genetic variance with a distinct internalizing (mood–anxiety–neuroticism) factor. Partitioned heritability indicated enrichment in several cortical and subcortical regions, and imputed genetically regulated gene expression in these regions was used to identify potential drug repositioning candidates. These results validate the biological coherence of the PTSD syndrome, inform its relationship to comorbid anxiety and depressive disorders and provide new considerations for treatment.

PTSD is a serious mental disorder that can occur after exposure to extreme, life-threatening stress[1,2]. Although 50–85% of Americans experience traumatic events over a lifetime, most do not develop PTSD—lifetime PTSD prevalence is approximately 7% (ref. [3]), suggesting differential resilience to stress and vulnerability to the disorder[4]. There is a substantial heritable basis for PTSD risk[5,6], and evidence from genome-wide association studies (GWAS) shows that PTSD, like other mental disorders[7], is highly polygenic[8–13]. PTSD symptoms vary widely among individuals, and the current Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5) definition permits up to 163,120 unique conformations for assembly of the disorder[14]. Given that this phenotypic heterogeneity may impede the detection of genetic risk factors[15], alternative phenotypes or subphenotypes (for example, re-experiencing (also known as intrusion) symptoms) that may reflect more biologically homogeneous entities have been examined[16].

The use of biobanks with relatively large numbers of PTSD cases gives the opportunity to provide unprecedented sample size and, importantly, uniformity of phenotypic and genotypic platforms[17]. This investigation was conducted within the US Veterans Affairs Million Veteran Program (MVP)[18] and included several PTSD phenotypic definitions: a validated, algorithmically defined case-control definition using data from the electronic health record (EHR), which was subsequently meta-analyzed with the case-control Psychiatric Genomics Consortium (PGC)-PTSD GWAS[13]; and quantitative trait definitions encompassing PTSD subdomains based on recent self-reported symptoms: re-experiencing (in an expanded sample from that previously reported[16]), avoidance, hyperarousal and a total index of recent symptom severity (PCL-Total). These analyses were conducted separately in veterans of EUR and AFR ancestry (and in transancestral meta-analyses)[19,20]. The heritability of each of these phenotypes, as well as phenotypic and genetic (r_g) correlations, were examined with the aim of determining coherence among them; r_g with other behavioral and health-related traits was also examined. Results for the phenotype with the largest single-nucleotide polymorphism (SNP) heritability estimate were used to characterize PTSD genomic architecture with partitioned heritability, and transcriptome-wide analyses[21] were utilized to identify genes regulated in the brain regions of greatest relevance. Genomic structural equation modeling was used to determine genetic relationships between PTSD and clinically comorbid phenotypes from the

[1]VA San Diego Healthcare System, Psychiatry Service, San Diego, CA, USA. [2]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [3]Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, USA. [4]VA Connecticut Healthcare System, Psychiatry Service, West Haven, CT, USA. [5]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [6]VA Boston Healthcare System, Massachusetts Veterans Epidemiology Research and Information Center, Boston, MA, USA. [7]Department of Psychiatry, Boston University School of Medicine, Boston, MA, USA. [8]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]Clinical Epidemiology Research Center, VA Connecticut Healthcare System, West Haven, CT, USA. [10]College of Medicine, University of Kentucky, Lexington, KY, USA. [11]Office of the Director, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD, USA. [12]Department of Medicine, Yale University School of Medicine, New Haven, CT, USA. [13]Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA. [14]Food and Drug Administration, Center for Drug Evaluation and Research, Silver Spring, MD, USA. [15]These authors contributed equally: Murray B. Stein, Joel Gelernter. [16]Deceased: Ronald S. Duman. *Lists of authors and their affiliations appear at the end of the paper. **A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: mstein@health.ucsd.edu; joel.gelernter@yale.edu

internalizing spectrum[22]: major depressive disorder, anxiety and neuroticism.

The aims of these analyses are to provide: (1) a large, uniformly phenotyped GWAS of PTSD in military veterans; (2) thorough exploration of subphenotypes; (3) replication of key associations in other datasets; (4) demonstration of the architecture of genetic association with other health-related phenotypes; (5) investigation of brain regions implicated; and (6) extension to possible drug targets. These aims were all accomplished with the overarching goal of deepening biological understanding to advance precision medicine for PTSD.

Results

GWAS of algorithmically defined case-control PTSD. We first performed GWAS of PTSD in American veterans of EUR and AFR ancestry, basing diagnosis on a validated EHR algorithm[23] that had excellent discriminative ability for lifetime PTSD cases versus controls as determined by chart review (0.90 sensitivity, 0.97 specificity, 0.87 positive predictive value and 0.90 negative predictive value), and substantial agreement with gold-standard, clinician-administered PTSD scale interview (90.2% agreement and $\kappa = 0.75$ (95% confidence interval (CI): 0.62, 0.88))[17]. GWAS analyses were carried out (on two tranches of data genotyped on the same array platform at two different times) on SNP dosages imputed from 1000 Genomes Phase 3, using logistic regression for case-control traits and linear regression for continuous traits in PLINK 2.0 (ref. [24]) and separately by ancestry, adjusting for age, sex and the first ten within-ancestry principal components. Meta-analysis by tranche (and later by ancestral group) was performed using METAL[25]. Combat exposure information was available for only a subset (51.2%) of the sample (Supplementary Table 1), and GWAS of that subset yielded no genome-wide significant (GWS) findings (Supplementary Table 2 shows findings at $P < 10^{-6}$). However, genetic correlation (r_g) between the categorical trait (that is, diagnosis of) PTSD in those combat exposed and in all subjects irrespective of combat exposure status was 0.969 (s.e. = 0.049, $P = 7.64 \times 10^{-89}$), and therefore results for the latter larger, more informative, sample are presented here.

The PTSD case-control GWAS for the EUR sample included 36,301 algorithmically defined probable PTSD cases and 178,107 controls. Considering linkage disequilibrium (LD)-independent loci ($r^2 > 0.1$), we identified three distinct GWS ($P < 5 \times 10^{-8}$) genomic risk loci (Fig. 1 (top) and Supplementary Table 3a): on Chr11:28707675, rs10767744 (minor allele frequencies (MAF) = 0.39, $P = 1.75 \times 10^{-10}$), proximity mapped to *METTL15*; on Chr7:70219946, rs137999048 (MAF = 0.047, $P = 1.03 \times 10^{-8}$), proximity mapped to *AUTS2*; and on Chr7:1855531, rs7680 (MAF = 0.14, $P = 4.17 \times 10^{-8}$), proximity mapped to *mitotic arrest deficient 1-like 1* (*MAD1L1*). Regional Manhattan plots for each region are presented in Supplementary Fig. 1a–c.

The GWAS for the AFR sample included 11,920 probable PTSD cases and 39,116 controls (Extended Data Fig. 1 and Supplementary Table 3b) and identified two distinct GWS loci, one on Chr3:1259951, rs4684090 (MAF = 0.04, $P = 3.59 \times 10^{-8}$) intronic to *CNTN6* and one on Chr20:6724577, rs112149412 (MAF = 0.02, $P = 3.19 \times 10^{-9}$) near *BMP2*. GWAS for the 48,221 cases and 217,223 controls in the transancestral analysis (meta-analysis of EUR and AFR samples) (Supplementary Table 3c) identified as GWS SNPs in two of the same regions found GWS in the EUR GWAS: a different lead SNP on Chr7:1959634 (rs137944087, an indel/deletion) in moderate LD with the variant identified in the EUR sample ($r^2 = 0.38$), and a different lead SNP on Chr11:28678870 (rs10767739) in LD with the variant identified in the EUR sample ($r^2 = 0.54$).

Meta-analysis of MVP and PGC-PTSD case-control GWAS. We next conducted meta-analyses of the EUR MVP and PGC-PTSD case-control GWAS[13] (Fig. 1 (bottom) and Supplementary Table 4a). The EUR meta-analysis yielded four distinct GWS loci, two of which were nearest to genes found to be GWS in the MVP case-control analysis (*MAD1L1* and *METTL15*), although with different lead SNPs: one new SNP (nearest to *LOC645949*) and one lead SNP closest to *PACRG*, a gene linked in a head-to-head arrangement and coregulated with *PARK2*—a gene found to be GWS in PGC. There were no GWS SNPs for the AFR MVP/PGC-PTSD meta-analysis, but two SNPs were GWS in the transancestral meta-analysis with lead SNPs closest to *PARK2* and *MAD1L1*, respectively (Supplementary Table 4b,c).

GWAS of PTSD symptom subphenotypes and total symptoms. The MVP surveys included the PTSD checklist for DSM-IV (PCL), a widely used, 17-item self-report measure of past-month PTSD symptoms covering the three DSM-IV symptom cluster criteria—re-experiencing, avoidance and hyperarousal—and a total symptom severity score (PCL-Total) as the sum of those three subphenotypes[26]. GWAS with these phenotypes in the EUR sample ($n = 186{,}689$ individuals) using linear regression identified multiple independent GWS SNPs, including some that were associated with PCL-Total as well as multiple subdomains, and others that were more strongly associated with specific subdomains (Table 1). Overlap in risk loci for the case-control and quantitative phenotypes in the EUR and AFR samples is shown in Fig. 2. Supplementary Table 5 shows PCL-Total GWAS results in the transancestral sample.

Fine-mapping and variant prioritization. For PCL-Total, we identified 15 GWS loci in the EUR population; for the case-control phenotype, we observed three loci in the EUR population and two in the AFR population. Each locus that included more than ten GWS SNPs was fine-mapped[27] to prioritize variants in each locus, defined as credible sets (Supplementary Data 1). Regions associated with PCL-Total scores had multiple variants with combined annotation-dependent depletion (CADD) score > 10 (that is, these variants were among the top 10% of pathogenic variants across the human genome)[28]. For example, in the region Chr3:49734229–50176259 associated with PCL-Total, there were four subregions with one or more exonic SNPs with CADD > 10. Fine-mapping results of causal variant identification in associated regions (CAVIAR)[27] and CADD scores are included in Supplementary Data 1.

To understand the biological effect of SNPs associated with PTSD phenotypes, we analyzed top SNPs (at suggestive threshold $P < 5 \times 10^{-6}$) for their distinct and overlapping distribution across the four subphenotypes. The top SNPs for each phenotype were LD pruned to obtain independent signals. We found 87 (hyperarousal), 49 (avoidance), 62 (re-experiencing) and 36 (PCL-Total) SNPs that were nonoverlapping or phenotype-specific (Supplementary Data 2). These nonoverlapping SNPs were assessed for their quantitative trait loci (QTL) protein associations (all tissues), DNA methylation (brain tissues) and splicing (brain tissues) from QTLbase[29]. Most QTL associations were observed for methylation expression and are shown as Venn diagrams for each phenotype (Supplementary Data 2); detailed tabular results are also given in Supplementary Data 2.

Replication of GWAS findings. We compared top SNP associations from the PTSD case-control and PCL-Total results against the largest available external PTSD dataset, from the PGC-PTSD[13]. For the EUR case-control phenotype, there was nominal replication for one of three SNPs: for rs7680*A nearest to *MAD1L1*, with a log(odds ratio(OR)) of −0.0712 (s.e. = 0.013, $P = 4.17 \times 10^{-8}$) in MVP and a log(OR) of −0.0639 (s.e. = 0.0215, $P = 0.00312$) in PGC-PTSD. For the EUR PCL-Total symptom scores, there were six of 15 possible nominal replications (Supplementary Table 6).

We applied a polygenic risk score (PRS) in EUR with MVP as the base and PGC as the target. The MVP case-control and MVP PCL-Total PRS explained approximately 0.4% ($P = 2.4 \times 10^{-74}$) and

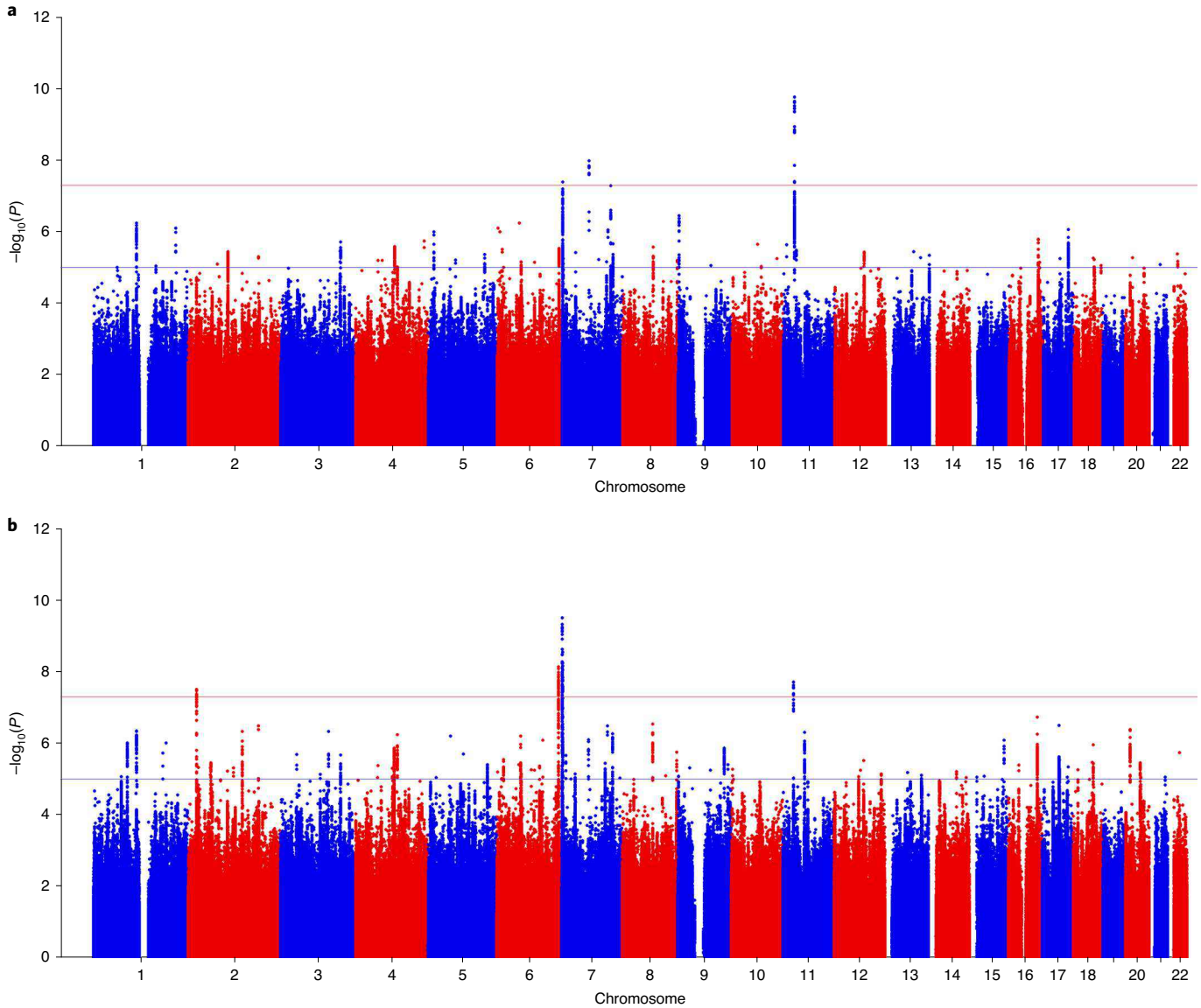


Fig. 1 | Manhattan plots for MVP case-control GWAS and MVP/PGC GWAS meta-analysis in EUR samples. a, MVP case-control GWAS. **b**, MVP/PGC GWAS meta-analysis. GWAS was performed using logistic regression, covarying for age, sex and the first ten principal components of ancestry. Meta-analysis was conducted with METAL[25] using the inverse variance weighting method. Bonferroni correction was used to correct for multiple comparisons; associations with $P < 5 \times 10^{-8}$ (indicated by the horizontal red bar) were considered to be GWS, and those with $P < 10^{-5}$ (indicated by the horizontal blue bar) are also shown.

0.7–0.8% of the variance ($P = 2.2 \times 10^{-134}$), respectively, in the PGC case-control phenotype at *P* value threshold (P_T) ≤ 0.05 (Extended Data Fig. 2). The low phenotypic variance explained is probably due to different characteristics of the MVP and PGC-PTSD cohorts: across three MVP hold-out PRS analyses we observed phenotypic variance explained ranging from 4 to 5.3% ($P < 6 \times 10^{-92}$; Supplementary Table 7). Evaluating the extent to which cross-ancestral PRS were useful, we found PRS biased by ancestry, with density plots of EUR and AFR PRS being substantially different (Extended Data Fig. 3).

SNP-based heritability estimates and genetic correlations across PTSD phenotypes and with other health-related traits. Figure 3 shows SNP-based heritability estimates (on the left) and the phenotypic (above the diagonal) and genetic (below the diagonal) correlations in EUR between the algorithmic case-control diagnosis and each of the four continuous PTSD symptoms (re-experiencing, avoidance, hyperarousal and their total; and the genetic correlations for the MVP/PGC case-control meta-analysis). Genetic correlations were consistently high ($r_g > 0.9$) across all PTSD traits, indicating that the traits investigated are all informative with respect to PTSD genetics. The PCL-Total quantitative trait (95% CI SNP heritability (SNP-h^2) = 0.08–0.10) has significantly higher SNP-based heritability than either the MVP case-control definition (95% CI SNP-h^2 = 0.05–0.07, $P_{difference} = 1.85 \times 10^{-4}$) or the MVP/PGC case-control meta-analysis (95% CI SNP-h^2 = 0.07–0.08, $P_{difference} = 5.83 \times 10^{-3}$), and significantly larger SNP heritability *z*-score (MVP PCL-total SNP-h^2 $z = 17.73$; MVP case-control SNP-h^2 $z = 11.62$; MVP/PGC SNP-h^2 $z = 14.80$).

In the EUR sample, we estimated genetic correlations (r_g) between PTSD case-control and PCL-Total scores and health-related traits available from UK Biobank and the PGC (Supplementary Table 8

Table 1 | GWS ($P < 5 \times 10^{-8}$) findings using linear regression with lead SNPs for EUR PCL-Total and subphenotype GWAS analyses ($n = 186{,}689$ individuals)

LD-independent lead SNP	Chr	Effect allele	β	P	INFO score	SNP location	Nearest gene
PCL-Total							
rs542933551	17	AAAAACAAAAC	0.4585	2.02×10^{-13}	0.95	43557054	*PLEKHM1*
rs10235664	7	C	−0.3667	1.82×10^{-11}	0.93	2086814	*MAD1L1*
rs35761884	1	C	−0.3076	3.46×10^{-10}	0.92	73787732	*LINC01360*
rs111488606	3	CA	0.3102	1.72×10^{-9}	0.83	49864924	*TRAIP*
rs13262595	8	G	−0.2823	2.20×10^{-9}	1.00	143316970	*TSNARE1*
rs2314662	19	C	−0.3614	3.78×10^{-9}	0.93	18702515	*C19orf60*
rs10171148	2	A	0.2811	5.87×10^{-9}	0.96	22466171	*LOC102723362*
rs62465629	7	C	−0.3929	6.30×10^{-9}	0.85	110153866	*IMMP2L*
rs1496246	11	G	0.2973	6.60×10^{-9}	0.90	133548061	*OPCML*
rs251350	5	C	−0.2538	1.03×10^{-8}	1.12	140225137	*PCDHA1*
rs11507683	9	T	0.4137	1.15×10^{-8}	0.96	140262424	*EXD3*
rs599550	18	A	0.3948	1.18×10^{-8}	0.95	53252388	*TCF4*
rs4364183	3	A	0.3043	1.22×10^{-8}	0.93	18809536	*SATB1-AS1*
rs62417832	6	T	0.2922	2.90×10^{-8}	1.00	88640221	*SPACA1*
rs111950471	5	TATTA	−0.2769	4.34×10^{-8}	0.98	107450098	*FBXL17*
Re-experiencing							
rs35371867	18	A	0.1006	1.24×10^{-10}	0.97	53193027	*TCF4*
rs2777888	3	G	0.0929	2.26×10^{-10}	0.98	49898000	*CAMKV*
rs10235664	7	C	−0.1055	4.66×10^{-10}	0.93	2086814	*MAD1L1*
rs242925	17	T	−0.0931	5.50×10^{-10}	0.94	43888866	*CRHR1*
rs139356208	11	CACAAAACAAA	−0.0897	9.63×10^{-9}	0.90	28631779	*RASEF*
rs1501485	1	G	−0.0839	1.22×10^{-8}	0.97	73995259	*LRRIQ3*
rs11773880	7	G	−0.0977	1.97×10^{-8}	0.93	106540171	*PIK3CG*
rs34177209	19	A	0.1205	2.34×10^{-8}	0.62	18474978	*PGPEP1*
rs10977193	9	A	−0.0934	4.17×10^{-8}	0.96	8542019	*PTPRD*
Avoidance							
rs55925547	17	C	0.1932	2.08×10^{-13}	0.98	43556807	*PLEKHM1*
rs199913382	17	C	0.1772	1.05×10^{-12}	0.98	44625866	*LRRC37A2*
rs35761884	1	C	−0.1388	9.72×10^{-11}	0.92	73787732	*LINC01360*
rs251350	5	C	−0.1192	8.15×10^{-10}	1.12	140225137	*PCDHA1*
rs4129585	8	C	−0.125	1.25×10^{-9}	1.00	143312933	*TSNARE1*
rs2314662	19	C	−0.1599	2.74×10^{-9}	0.93	18702515	*C19orf60*
rs62465629	7	C	−0.175	3.54×10^{-9}	0.85	110153866	*IMMP2L*
rs62417832	6	T	0.1335	7.04×10^{-9}	1.00	88640221	*SPACA1*
rs11507683	9	T	0.1834	7.74×10^{-9}	0.96	140262424	*EXD3*
rs10171148	2	A	0.1211	1.07×10^{-8}	0.96	22466171	*LOC102723362*
rs10235664	7	C	−0.1337	2.17×10^{-8}	0.93	2086814	*MAD1L1*
rs1496246	11	G	0.1234	3.66×10^{-8}	0.90	133548061	*OPCML*
Hyperarousal							
rs377112142	17	CT	0.1323	3.06×10^{-13}	0.84	43663455	*MAPK8IP1P2*
rs55789728	7	G	−0.1303	4.62×10^{-13}	0.93	2107649	*MAD1L1*
rs576430065	9	CA	−0.1206	1.67×10^{-11}	0.78	96373697	*PHF2*
rs140288713	17	A	0.1286	3.11×10^{-11}	0.90	44690708	*NSFP1*
rs1496246	11	G	0.1037	1.77×10^{-10}	0.90	133548061	*OPCML*
rs547649546	3	CA	−0.0937	1.59×10^{-9}	0.91	49789921	*IP6K1*
rs2887882	1	T	−0.1118	1.89×10^{-9}	0.98	113170389	*CAPZA1*

Continued

Table 1 | GWS ($P < 5 \times 10^{-8}$) findings using linear regression with lead SNPs for EUR PCL-Total and subphenotype GWAS analyses ($n = 186,689$ individuals) (continued)

LD-independent lead SNP	Chr	Effect allele	β	P	INFO score	SNP location	Nearest gene
rs7519147	1	T	-0.0906	1.90×10^{-9}	0.96	73994416	*LRRIQ3*
rs13032994	2	C	-0.0968	3.73×10^{-9}	1.00	52709559	*NRXN1*
rs113341106	7	GC	0.0923	3.82×10^{-9}	0.93	114039998	*FOXP2*
rs12420134	11	G	0.1229	6.45×10^{-9}	0.87	16260861	*SOX6*
rs17209774	9	C	-0.0907	7.97×10^{-9}	0.97	4145163	*GLIS3*
rs60958094	14	T	0.0961	1.99×10^{-8}	0.81	54711168	*CDKN3*
rs4129585	8	C	-0.0835	2.07×10^{-8}	1.00	143312933	*TSNARE1*
rs549326362	5	T	-0.0884	4.46×10^{-8}	0.94	107444481	*FBXL17*

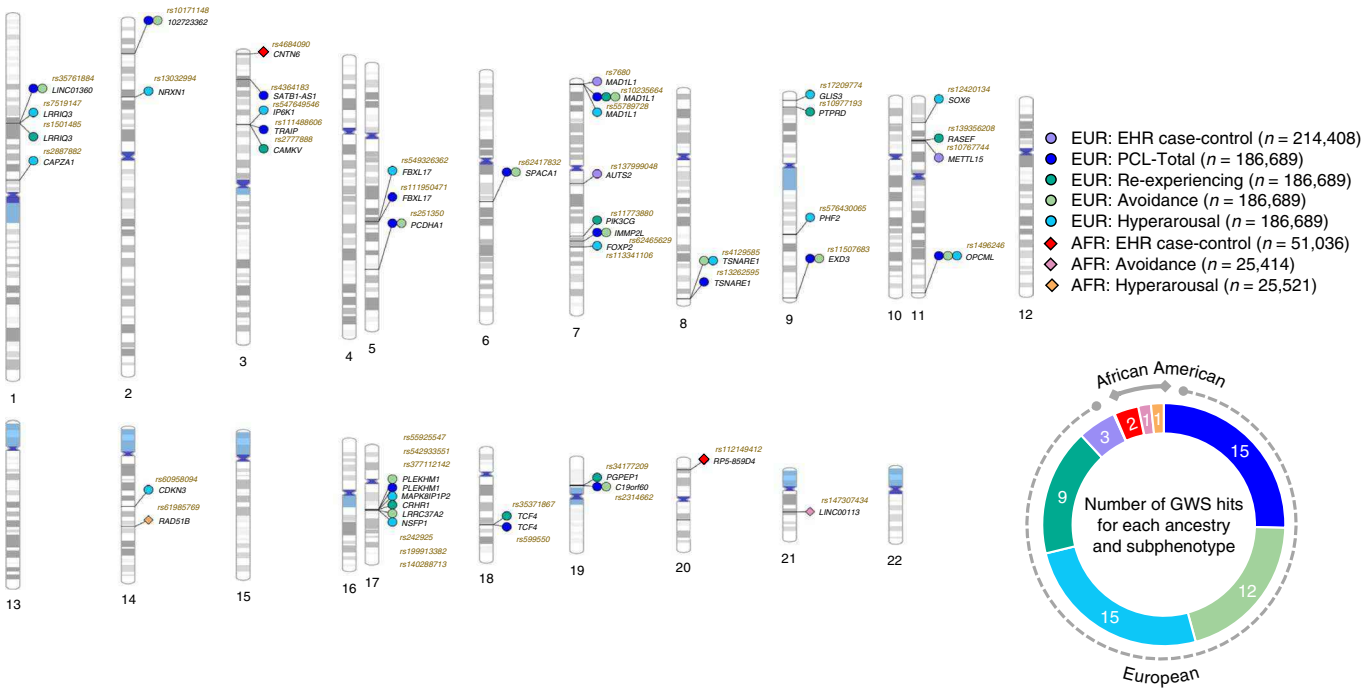


Fig. 2 | GWS findings, by ancestry and PTSD phenotype. Results are shown for EUR (circles) and AFR (diamonds) ancestry, for PTSD EHR case-control, PCL-Total and PTSD subcomponents (avoidance, hyperarousal and re-experiencing). There were no GWS results for AFR PTSD case-control and re-experiencing traits. LD-independent SNPs for each phenotype and the nearest gene are labeled. The donut chart summarizes the number of hits for each phenotype in the two ancestral populations. The genes labeled are significant following regression testing for a two-sided *P* value with applied Bonferroni threshold for multiple testing ($P < 5 \times 10^{-8}$).

shows r_g for all traits with h^2 *z*-score of 4 or more). The many significant genetic correlations with both PTSD traits include positive r_g with major depression, neuroticism and related symptoms, and negative r_g with educational attainment and cognitive performance (Fig. 4). Although the magnitudes of r_g observed with PCL-Total and case-control PTSD were highly correlated (Spearman's rho = 0.970, $P = 2.20 \times 10^{-16}$), ten phenotypes exhibited significantly greater r_g with PCL-Total relative to case-control PTSD (Fig. 4a).

Taken together, the higher heritability, the greater magnitude of the heritability *z*-score (indicative of a stronger polygenic signal) and the higher value of r_g with other health-related traits confirm that PCL-Total is similar to, but more informative than, the case-control definition (for either MVP alone or the MVP/PGC meta-analysis). Accordingly, all subsequent post-GWAS analyses are based on the more powerful PCL-Total quantitative trait dataset in the EUR sample.

Genomic relationship between PTSD and other mental disorders. We used multitrait conditional and joint analysis (mtCOJO)[30] to address the genetic relationship between PTSD and other major mental disorders in two ways. First, we conditioned PTSD PCL-Total on a single mental disorder; then, we conditioned PTSD PCL-Total on all eight mental disorders simultaneously: autism spectrum disorder, major depression, anorexia nervosa, anxiety (case-control), alcohol dependence, schizophrenia, bipolar disorder and attention deficit hyperactivity disorder[31–38]. The result of this analysis is treated as genetic signal attributable to PTSD in the absence of shared genetic liabilities of other mental disorders. PCL-Total remained highly genetically correlated with unconditioned GWAS when conditioned on genetically correlated psychiatric disorders, both independently (that is, PTSD PCL-Total conditioned on major depreesive disorder) and simultaneously (that is, PTSD PCL-Total conditioned on all eight mental disorders; Fig. 5). Conditioning

	MVP/PGC case-control	MVP case-control	PCL-Total	Re-exp	Avoid	Hyper
MVP case-control h_2 = 6.4%	0.974 0.965–0.982		0.86 0.857–0.860 *n* = 111,362	0.83 0.831–0.835 *n* = 91,879	0.82 0.824–0.826 *n* = 110,739	0.8 0.796–0.800 *n* = 112,133
PCL-Total h_2 = 9.2%	0.969 0.943–0.994	0.959 0.903–1.014		0.92 0.923–0.925 *n* = 141,076	0.96 0.964–0.965 *n* = 160,504	0.93 0.932–0.934 *n* = 162,348
Re-exp h_2 = 9.3%	0.971 0.945–0.996	0.977 0.903–1.014	0.973 0.966–0.979		0.85 0.849–0.852 *n* = 130,341	0.80 0.801–0.805 *n* = 145,990
Avoid h_2 = 9.3%	0.93 0.902–0.959	0.915 0.856–0.974	0.984 0.98–0.988	0.931 0.916–0.946		0.85 0.849–0.852 *n* = 159,002
Hyper h_2 = 10.1%	0.944 0.919–0.97	0.953 0.896–1.009	0.979 0.972–0.987	0.935 0.919–0.951	0.943 0.929–0.958	

Fig. 3 | Phenotypic and genetic correlations between case-control, PCL-Total and subscale scores. Shown are correlation point estimates (top of box), 95% CIs (middle) and *n* (bottom; sample size); phenotypic: above black-boxed diagonal; genetic: below diagonal. SNP heritability (h^2) is shown in the left-hand column. For phenotypic correlations, those for case-control are point-biserial while all others are Pearson correlations. Re-exp, re-experiencing; avoid, avoidance; hyper, hyperarousal.

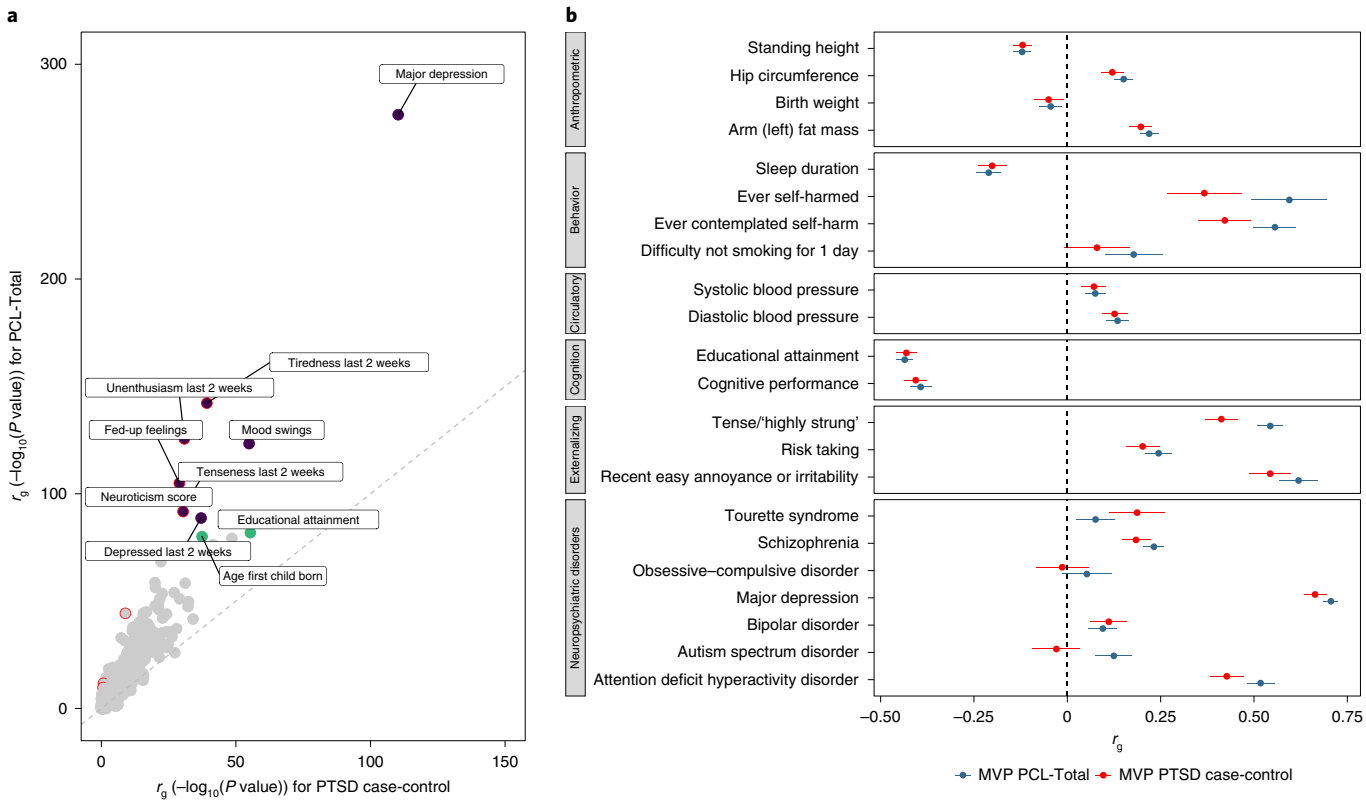


Fig. 4 | LDSC genetic correlation analyses in EUR showing traits from UK Biobank and PGC psychiatric disorders. a, Comparison of r_g between PTSD case-control definition and PCL-Total. The dashed gray diagonal indicates perfect linearity between PTSD case-control and PCL-Total genetic correlates. The top ten genetic correlates of PCL-Total are labeled, with purple and green data points indicating positive and negative r_g, respectively. Data points circled in red indicate significant difference in r_g between PTSD and PCL-Total. **b**, Plot showing a wide range of phenotypes and their r_g values; the vertical dashed line indicates $r_g = 0$.

on all eight mental disorder traits significantly reduced the observed scale of SNP heritability (h^2) of PCL-Total (PCL-Total original $h^2 = 9.21\%$, $P = 1.39 \times 10^{-67}$; PCL-Total conditioned $h^2 = 4.11\%$, $P = 2.61 \times 10^{-52}$) relative to unconditioned GWAS ($P_{difference} = 1.52 \times 10^{-13}$), but this reduction in heritability did not significantly alter associations with biological pathways or tissues

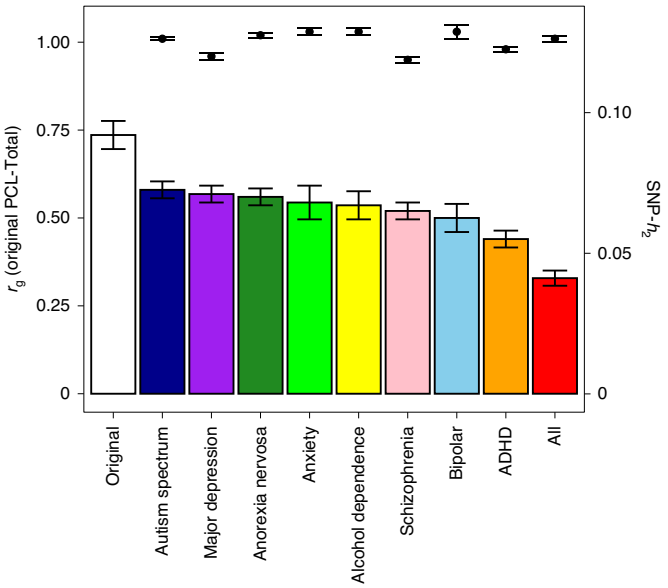


Fig. 5 | Genetic relationship between PCL-Total and other mental health phenotypes. PCL-Total SNP-h^2 error bars and standard error (right *y* axis) and r_g data points and standard error (left *y* axis) relative to original PCL-Total ($n = 186,689$) after conditioning with each mental health phenotype on the *x* axis. ADHD, attention deficit hyperactivity disorder.

associated with genetic risk for PTSD, as evidenced by the linear relationships between tissue and pathway enrichment effects (Extended Data Fig. 4).

Genomic structural equation modeling. Genomic structural equation models were analyzed to answer two question: (1) do PTSD subdomains (hyperarousal, re-experiencing and avoidance) load onto one latent factor? (2) Does latent factor architecture and subdomain loading change in the presence of PTSD genetic and phenotypic correlates—major depressive disorder, anxiety and neuroticism? These traits—all part of the internalizing spectrum[22,39]—are highly phenotypically and genetically correlated with PTSD.

The three PTSD phenotypic subdomains loaded onto a single latent common factor (Supplementary Fig. 2). There were no significant differences in loading values between these PTSD subdomains, suggesting roughly equal contribution of all three to the common factor (comparative fit index (CFI) = 0.996). Next, we included PTSD genetic and phenotypic correlates from internalizing disorders—anxiety, neuroticism and major depressive disorder (all from PGC). Genomic exploratory factor analysis identified a two-factor model as best suited to represent the six phenotypes (that is, PTSD subdomains and the three internalizing measures). In genomic confirmatory factor analysis of the two-factor model, PTSD subdomains independently loaded onto factor 1 while the PTSD correlates loaded onto a second factor (CFI = 0.999) (Fig. 6). The PTSD subdomain hyperarousal loads onto both factors (loading onto factor 1 = 0.90 ± 0.05; loading onto factor 2 = 0.10 ± 0.04; correlation between factors 1 and 2 = 0.72 ± 0.03), indicating that this subdomain has a genetic correlation with the internalizing psychopathologies that is not shared by the other PTSD subdomains.

Partitioned heritability of PCL-Total. Partitioning heritability of PCL-Total in EUR revealed enrichment of SNPs (by 1.28- to 1.39-fold) associated with four genotype–tissue expression (GTEx) cortical tissue types: cortex, frontal cortex (BA9), anterior cingulate cortex (BA24) and nucleus accumbens (false discovery rate (FDR) $q < 0.05$; Supplementary Table 9). Intronic regions showed 1.29-fold enrichment (FDR $q < 0.05$). Cell-type partitioning analyses support SNP-h^2 enrichment of the frontal cortex (BA9) gene sets (Tau-C = 3.42×10^{-9}, $P = 0.002$) above other annotations in the model, and frontal cortex (BA9), anterior cingulate cortex (BA24) and multiple basal ganglia (putamen, caudate and nucleus accumbens) gene expression profiles (Tau-C ranging from 1.02×10^{-9} to 3.43×10^{-9}, FDR $q < 0.05$) above that of all other genomic annotations. These tissues were prioritized when considering transcriptome-wide association results, to constrain interpretation of those results to the most pertinent and evidence-driven tissues[40].

Enrichment in biological tissues using transcriptome-wide analysis and colocalization. PrediXcan-S[41] was used to correlate imputed tissue-specific, genetically regulated gene expression determined by association with reference transcriptome datasets with PCL-Total results. We observed significant negative correlation with predicted expression of the protein product of the pseudogene *LRRC37A4P* in amygdala, substantia nigra, putamen, frontal and anterior cingulate cortex, adrenal gland and whole-blood tissues. Also noted were significant positive correlations with predicted expression of *corticotropin-releasing hormone receptor 1* (*CRHR1*) in amygdala, hippocampus, frontal and anterior cingulate cortex, adrenal and whole blood (although negative correlation was seen for nucleus accumbens); significant positive correlation with predicted expression of *PLEKHM1*, *ARL17A*, *LRRC37A2* and *DND1P1* (all of which are colocalized on 17q21.31) in multiple brain regions, including amygdala, anterior cingulate cortex and basal ganglia; and significant negative correlation with predicted expression of *RBM6* in frontal cortex, hippocampus, nucleus accumbens, adrenal and whole blood. The complete list of PrediXcan-S results is available in Supplementary Table 10. The significant genes for 13 brain tissues were then tested for shared causal loci. The coloc method[42] reports posterior probability for a pair of traits under the hypothesis (H_4) that traits are associated and share a single causal variant. The genetically regulated transcriptomic profiles of *ARL17A*, *LRRC37A2*, *RNF123*, *FAM212A* and *PLEKHM1* showed high probability (≥90%) of a shared causal locus (coloc H_4) with PCL-Total across multiple brain regions. *CRHR1* probability was highest (85%) for hippocampus tissue expression (Supplementary Data 2).

Drug repositioning analyses. We selected genes significantly associated with PCL-Total in the PrediXcan-S analyses and, as recommended[40], prioritized those genes with predicted expression regulation in at least one of the four tissues identified by LD score regression (LDSC) partitioned heritability analyses: cortex, frontal cortex, anterior cingulate and nucleus accumbens (Fig. 7).

We imported this list of eight genes (*ARHGAP27*, *ARL17A*, *CRHR1*, *DND1P1*, *LRRC37A2*, *LRRC37A4P*, *PLEKHM1* and *RBM6*) into the Drug Gene Interaction Database v.3.0 (dgidb.genome.wustl.edu)[43] to identify interactions with available drug treatments that might indicate potential new drug strategies for PTSD. Drug repositioning analysis was also carried out in the Connectivity Map (CMap) database (https://www.broadinstitute.org/connectivity-map-cmap) and PHAROS (https://pharos.nih.gov) for the same set of eight genes[44].

No currently druggable targets were identified for *ARHGAP27*, *ARL17A*, *DND1P1*, *LRRC37A2*, *LRRC37A4P* or *RBM6*. *CRHR1* was identified in all databases as a potential drug target with experimental medications available. Given the positive association between PTSD symptoms and imputed *CRHR1* expression in multiple brain regions (with the exception of nucleus accumbens) seen in our dataset, a *CRHR1* antagonist is hypothesized to be potentially therapeutic. Another gene, *PLEKHM1*, which was significantly associated with imputed increased expression and colocalized in caudate and nucleus accumbens, was considered by CMap as highly likely to

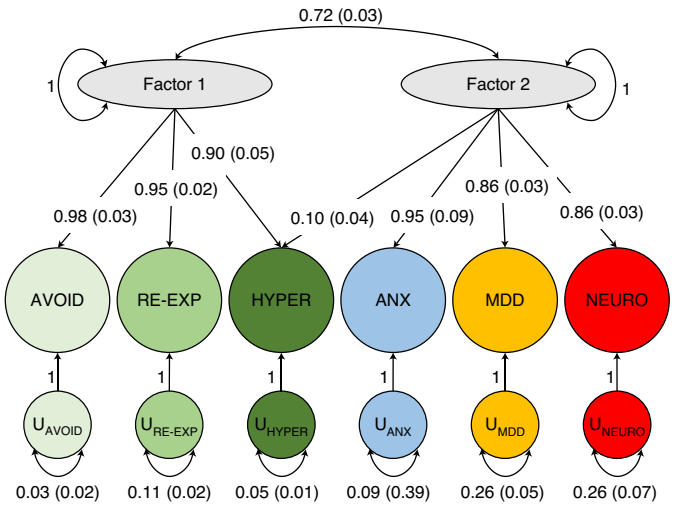


Fig. 6 | GenomicSEM model with confirmatory factor analysis indicating two correlated factors. The first factor consists of PTSD symptoms and the second consists of anxiety, major depressive disorder and neuroticism. CFI = 0.999 (typically interpreted as 0.90–0.95, indicating marginal fit). Each trait loads onto at least one factor, with standard errors in parentheses and residual variance parameters indicated by each U bubble. AVOID, avoidance; RE-EXP, re-experiencing; HYPER, hyperactivity; ANX, anxiety; MDD, major depressive disorder; NEURO, neuroticism.

share biological effects with several classes of drug, including dopamine receptor antagonists, acetylcholine receptor antagonists and alpha-2 adrenergic receptor and angiotensin receptor antagonists, all of which would be predicted to reduce expression and be associated with a reduction in PTSD symptoms.

Discussion

The past decade has seen a proliferation in the use and usefulness of GWAS, with the prediction—and, to date, the experience—that continued sample size growth will result in even richer findings[45]. The field of psychiatric genomics has capitalized on GWAS, with substantial gains made in the understanding of serious mental disorders such as schizophrenia, major depression, bipolar disorder[7,46] and their interrelatedness[47]. We present here a large, uniformly phenotyped and genotyped case-control GWAS of PTSD in military veterans. We augment this analysis with the GWAS of a quantitative trait corresponding to symptom severity, which proved more statistically powerful than the case-control analysis even when our case-control GWAS was meta-analyzed with the next largest PTSD case-control GWAS available, from the PGC[13].

These analyses revealed several GWS associations with PTSD visible at the case-control level, and numerous GWS associations with various dimensions of symptom severity. When combined with imputed genetically regulated expression results and enrichment analyses, these results help to illuminate the neurobiology of PTSD and begin to uncover new avenues for therapeutic development.

This study directly compares the heritability of binary (diagnostic) and continuous (symptom-based) phenotypes for PTSD. Although PTSD symptoms can have a very diverse phenotypic presentation[14], we show here that their genetic overlap is very high ($r_g > 0.9$). This is an important insight into the biology of PTSD. The quantitative (PCL-Total) trait—which reflects the most information—was the most heritable and therefore the most informative for biological inference. Partitioned heritability analyses of that trait indicated overrepresentation of SNPs in frontal (BA9) and anterior cingulate cortex (BA24), consistent with prevailing neural circuit theories of PTSD pathophysiology[2] that emphasize hypofunction of these regions and their connections with the limbic cortex in the regulation of emotion and extinction of fear memories[48,49]. However, these analyses also pointed to the nucleus accumbens—an important component of the reward system—as being involved in PTSD symptoms. These results suggest that more extensive study of the nucleus accumbens and reward systems in PTSD may shed further light on aspects of the syndrome (for example, its strong association with alcohol dependence)[50,51] that are currently not well understood.

Several genes—most notably *MAD1L1*—were repeatedly implicated across the various conceptualizations of the PTSD phenotype. The variants in *MAD1L1* also show QTL associations with DNA methylation and splicing. *MAD1L1*, widely expressed in all tissues and thought to play a role in cell cycle control, has emerged as being GWS associated with at least two other major mental disorders, schizophrenia[31] and bipolar disorder[38]—both of which were excluded among participants in this study but have strong genetic correlations with PTSD in MVP and other cohorts[13]. These observations, and the recent finding of GWS association with anxiety[52], suggest that *MAD1L1* may be a general risk factor for psychopathology, possibly contributing to the *p* factor thought to underlie many serious mental disorders[53].

Several other genes were discovered to be associated with PTSD and replicated in the largest available independent PTSD-informative dataset, the PGC-PTSD GWAS[13]. Included among these were *TSNARE1* (*T-SNARE Domain Containing 1*) and *EXD3* (*Exonuclease 3′–5′ Domain Containing 3*). *TSNARE1*, the product of which is involved in intracellular protein transport, has been associated with risk taking[54], which may predispose to PTSD through increasing the likelihood of exposure to traumatic events; twin studies suggest that risk for exposure to traumatic events is partially heritable[5]. *EXD3*, the product of which is involved in nucleic acid binding, has been associated with mathematical[55] and other cognitive abilities, which have been found in our study and others to be genetically negatively correlated with PTSD and mediated by socioeconomic status[56]. The MVP/PGC case-control meta-analyses also identified associations with *PARK2* and *PACRG*, both of which are associated with susceptibility to both leprosy and intracellular pathogens[57]. It remains to be determined to what extent these associations reflect systems or processes that underlie PTSD pathophysiology, but we now have gene candidates discovered and replicated through unbiased searches that can be further examined in relation to their putative biological relationships to PTSD and other stress- and anxiety-related conditions. Supplementary Note includes a discussion of fine-mapping, functional annotation and CADD scores.

Analyses adjusting for genetic signals attributable to other major psychiatric disorders verified shared heritability with these other disorders while simultaneously confirming residual, distinct heritability for PTSD. The high value of r_g between PTSD symptom subdomains, which do not include overlapping items, supports the coherence of PTSD as a diagnostic construct from a biological perspective: that is, the same genetic predisposition underlies different symptoms previously identified as syndromic. Genomic structural equation modeling recapitulated genetic and phenotypic correlations between PTSD subdomains, suggesting that each PTSD subdomain is largely explained by the same genetic architectures. Our model also suggests that, whereas PTSD symptoms constitute a genetically distinct and cohesive module, hyperarousal may be a relevant subdomain linking the genetic and phenotypic relationships between PTSD, anxiety, major depressive disorder and neuroticism.

CRHR1 is in a large LD block on chromosome 17, making it difficult to discern its association with PTSD apart from other genes in that LD block. In our previous study of intrusive re-experiencing symptoms in MVP, we supported *CHRH1* as the gene with the

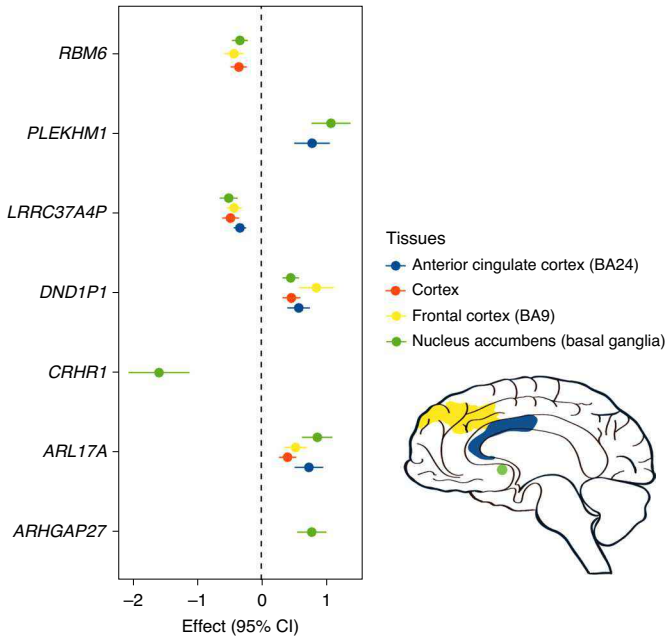


Fig. 7 | Genetically regulated transcriptomic changes with PCL-Total ($n = 186,689$). Effect (beta) and 95% CI values are shown on the *x* axis for each association across the four brain regions identified by partitioned heritability analyses using LDSC.

strongest association using transancestral meta-analysis[16]. We now provide additional biological evidence that *CRHR1* may be causally related to PTSD. PrediXcan-S analyses pointed to increased expression of *CRHR1* in amygdala, hippocampus (the structure with the highest colocalization probability), frontal cortex and anterior cingulate, regions repeatedly implicated as structurally or functionally abnormal in PTSD[2]. These results must be replicated and extended to other brain regions such as ventromedial prefrontal cortex, shown to be integral to fear learning and extinction[58], processes hypothesized to be central to PTSD onset and recovery, respectively[2,59]. In concert with strong preclinical and clinical priors for the involvement of corticotropin-releasing hormone in stress-related disorders[60], these observations position drugs that influence *CRHR1* as strong therapeutic candidates for PTSD and related conditions. Whereas a placebo-controlled trial of a CRHR1 antagonist in 128 women with PTSD produced unimpressive results[61], our findings (albeit predominantly in men) suggest that there are potential unfulfilled opportunities with *CRHR1* antagonists for PTSD that should be further explored, taking into account individual variation in *CRHR1*—including epigenetic variation[62]—as a source of differential antagonist efficacy, in keeping with the march toward precision psychiatry[63]. Furthermore, our unexpected finding of a negative association between PTSD symptom severity and predicted *CRHR1* expression in nucleus accumbens—which suggests that an agonist might be therapeutic—requires further investigation.

Our findings also tentatively support consideration of several drug classes as therapeutic repurposing candidates for PTSD. For example, acetylcholine receptor antagonists could be considered given their association in CMap with *PLEKHM1*. In a recent rodent study, the muscarinic receptor antagonist, scopolamine, augmented extinction in conjunction with exposure[64] (although other studies suggest that positive allosteric modulation of M1 muscarinic activity enhances contextual fear conditioning)[65]. These results together suggest that a therapeutic role for cholinergic modulation in PTSD and other fear-related conditions, possibly in concert with exposure therapy, should be investigated. Angiotensin receptor antagonists, also identified as drug candidates through CMap, have a strong preclinical rationale for use in PTSD[66–68] and are, in fact, currently undergoing testing in a randomized, placebo-controlled trial of losartan for PTSD (ClinicalTrials.gov Identifier: NCT02709018).

Our study has limitations. It is not currently known whether genetic risk for PTSD differs by trauma type (for example, combat exposure versus civilian trauma exposure) or developmental timing (for example, childhood maltreatment versus adult assault). Such differences could possibly underlie clinically and biologically important heterogeneity[69]. Studies of even larger sample size (which MVP will attain in the coming years) and greater granularity with regard to types and chronology of trauma exposure will be needed to address these questions. It is also important to note that the PCL is a state, not a trait measure, and therefore reflects current—but not necessarily worst-ever lifetime—severity. Our study also reports on a large AFR-ancestry sample, which we leveraged by inclusion of those individuals in our transancestral meta-analyses, but we relied, out of necessity, on the EUR ancestry sample for the post-GWAS analyses. We found, as might have been anticipated given previous work[70], that PRS derived in the European sample did not predict well into the AFR sample. Nevertheless, we aspire to using new tools in the future to make better use of the ancestral diversity in MVP[20].

We used transcriptome-wide association approaches to inform our drug repurposing inquiries. As recommended[40], we attempted to limit tissue biases inherent to these approaches by constraining our sphere of interest to brain regions that were associated with PTSD severity through our partitioned heritability analyses. Nonetheless, the drug repurposing propositions, while hypothesis generating and intriguing, are just that. They are one piece of information that might increase interest in testing the proposed drug classes in patients with PTSD; they must be buttressed by additional preclinical models, postmortem PTSD brain studies[71] and complementary bioinformatic approaches[72] supporting their use, as well as by serious consideration of their safety in this population. We also remind readers that the present analyses rested solely on GWAS, thereby limiting inquiry to common genetic variants (to MAF = 0.01, which still capture significant heritable variance) and that roles for rare variants and structural variation should also be explored. Epigenetic factors almost certainly also play a role in a disorder such as PTSD[10,73], which has traumatic stress as its precursor. Many other functional genomics tools can, and should, be brought to bear on the study of PTSD, expanding the scope of inquiry to encompass a holistic, integrative functional genomic analysis[74] of this common, serious and yet still poorly understood and inadequately treated neuropsychiatric disorder.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00767-x.

Received: 4 March 2020; Accepted: 15 December 2020;
Published online: 28 January 2021

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn (American Psychiatric Publishing, 2013).
2. Shalev, A., Liberzon, I. & Marmar, C. Post-traumatic stress disorder. *N. Engl. J. Med.* **376**, 2459–2469 (2017).
3. Kessler, R. C. & Wang, P. S. The descriptive epidemiology of commonly occurring mental disorders in the United States. *Annu. Rev. Public Health* **29**, 115–129 (2008).

4. Atwoli, L., Stein, D. J., Koenen, K. C. & McLaughlin, K. A. Epidemiology of posttraumatic stress disorder: prevalence, correlates and consequences. *Curr. Opin. Psychiatry* **28**, 307–311 (2015).
5. Stein, M. B., Jang, K. L., Taylor, S., Vernon, P. A. & Livesley, W. J. Genetic and environmental influences on trauma exposure and posttraumatic stress disorder symptoms: a twin study. *Am. J. Psychiatry* **159**, 1675–1681 (2002).
6. Wolf, E. J. et al. A classical twin study of PTSD symptoms and resilience: evidence for a single spectrum of vulnerability to traumatic stress. *Depress. Anxiety* **35**, 132–139 (2018).
7. Sullivan, P. F. et al. Psychiatric genomics: an update and an agenda. *Am. J. Psychiatry* **175**, 15–27 (2018).
8. Stein, M. B. et al. Genome-wide association studies of posttraumatic stress disorder in 2 cohorts of US Army soldiers. *JAMA Psychiatry* **73**, 695–704 (2016).
9. Duncan, L. E. et al. Largest GWAS of PTSD (*N*=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol. Psychiatry* **23**, 666–673 (2018).
10. Daskalakis, N. P., Rijal, C. M., King, C., Huckins, L. M. & Ressler, K. J. Recent genetics and epigenetics approaches to PTSD. *Curr. Psychiatry Rep.* **20**, 30 (2018).
11. Benjet, C. et al. The epidemiology of traumatic event exposure worldwide: results from the World Mental Health Survey Consortium. *Psychol. Med.* **46**, 327–343 (2016).
12. Xie, P. et al. Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biol. Psychiatry* **74**, 656–663 (2013).
13. Nievergelt, C. M. et al. International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nat. Commun.* **10**, 4558 (2019).
14. Galatzer-Levy, I. R. & Bryant, R. A. 636,120 ways to have posttraumatic stress disorder. *Perspect. Psychol. Sci.* **8**, 651–662 (2013).
15. Stein, M. B. Genomics of posttraumatic stress disorder: sequencing stress and modeling misfortune. *Biol. Psychiatry* **83**, 795–796 (2018).
16. Gelernter, J. et al. Genome-wide association study of post-traumatic stress disorder reexperiencing symptoms in >165,000 US veterans. *Nat. Neurosci.* **22**, 1394–1401 (2019).
17. Radhakrishnan, K. et al. Genomics of posttraumatic stress disorder in veterans: methods and rationale for Veterans Affairs Cooperative Study #575B. *Int. J. Methods Psychiatr. Res.* **28**, e1767 (2019).
18. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
19. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
20. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
21. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
22. Genetics of Personality Consortiumet al. Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* **72**, 642–650 (2015).
23. Harrington, K. M. et al. Validation of an electronic medical record-based algorithm for identifying posttraumatic stress disorder in U.S. veterans. *J. Trauma Stress* **32**, 226–237 (2019).
24. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
25. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
26. Blanchard, E. B., Jones-Alexander, J., Buckley, T. C. & Forneris, C. A. Psychometric properties of the PTSD Checklist (PCL). *Behav. Res. Ther.* **34**, 669–673 (1996).
27. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
28. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
29. Zheng, Z. et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res.* **48**, D983–D991 (2020).
30. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
31. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
32. Otowa, T. et al. Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry* **21**, 1391–1399 (2016).
33. Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
34. Walters, R. K. et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).
35. Martin, J. et al. A genetic investigation of sex bias in the prevalence of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **83**, 1044–1053 (2018).
36. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
37. Duncan, L. et al. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *Am. J. Psychiatry* **174**, 850–858 (2017).
38. Stahl, E. et al. Genomewide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
39. Kotov, R. et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
40. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
41. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
42. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
43. Cotto, K. C. et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.* **46**, D1068–D1073 (2018).
44. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
45. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
46. Smoller, J. W. Psychiatric genetics begins to find its footing. *Am. J. Psychiatry* **176**, 609–614 (2019).
47. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).
48. Dunsmoor, J. E., Niv, Y., Daw, N. & Phelps, E. A. Rethinking extinction. *Neuron* **88**, 47–63 (2015).
49. Phelps, E. A. & Hofmann, S. G. Memory editing from science fiction to clinical practice. *Nature* **572**, 43–50 (2019).
50. Suh, J. & Ressler, K. J. Common biological mechanisms of alcohol use disorder and post-traumatic stress disorder. *Alcohol Res.* **39**, 131–145 (2018).
51. Smith, N. D. L. & Cottler, L. B. The epidemiology of post-traumatic stress disorder and alcohol use disorder. *Alcohol Res.* **39**, 113–120 (2018).
52. Levey, D. F. et al. Reproducible genetic risk loci for anxiety: results from approximately 200,000 participants in the Million Veteran Program. *Am. J. Psychiatry* **177**, 223–232 (2020).
53. Selzam, S., Coleman, J. R. I., Caspi, A., Moffitt, T. E. & Plomin, R. A polygenic *p* factor for major psychiatric disorders. *Transl. Psychiatry* **8**, 205 (2018).
54. Karlsson Linner, R. et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).
55. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
56. Polimanti, R. et al. Association of economic status and educational attainment with posttraumatic stress disorder: a Mendelian randomization study. *JAMA Netw. Open* **2**, e193447 (2019).
57. Meschede, J. et al. The parkin-coregulated gene product PACRG promotes TNF signaling by stabilizing LUBAC. *Sci. Signal.* **13**, eaav1256 (2020).
58. Dunsmoor, J. E. et al. Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. *J. Neurosci.* **39**, 3264–3276 (2019).
59. Maddox, S. A., Hartmann, J., Ross, R. A. & Ressler, K. J. Deconstructing the gestalt: mechanisms of fear, threat, and trauma memory encoding. *Neuron* **102**, 60–74 (2019).
60. Chrousos, G. P. & Zoumakis, E. Milestones in CRH research. *Curr. Mol. Pharmacol.* **10**, 259–263 (2017).
61. Dunlop, B. W. et al. Corticotropin-releasing factor receptor 1 antagonism is ineffective for women with posttraumatic stress disorder. *Biol. Psychiatry* **82**, 866–874 (2017).
62. Pape, J. C. et al. DNA methylation levels are associated with CRF1 receptor antagonist treatment outcome in women with post-traumatic stress disorder. *Clin. Epigenetics* **10**, 136 (2018).
63. Stein, M. B. & Smoller, J. W. Precision psychiatry – will genomic medicine lead the way? *JAMA Psychiatry* **75**, 663–664 (2018).
64. Craske, M. G., Fanselow, M., Treanor, M. & Bystritksy, A. Cholinergic modulation of exposure disrupts hippocampal processes and augments extinction: proof-of-concept study with social anxiety disorder. *Biol. Psychiatry* **86**, 703–711 (2019).

65. Maksymetz, J. et al. M1 muscarinic receptors modulate fear-related inputs to the prefrontal cortex: implications for novel treatments of posttraumatic stress disorder. *Biol. Psychiatry* **85**, 989–1000 (2019).
66. Marvar, P. J. et al. Angiotensin type 1 receptor inhibition enhances the extinction of fear memory. *Biol. Psychiatry* **75**, 864–872 (2014).
67. Shekhar, A. Angiotensin type 1 receptor antagonists—a novel approach to augmenting posttraumatic stress disorder and phobia therapies? *Biol. Psychiatry* **75**, 836–837 (2014).
68. Pulcu, E. et al. The effects of the angiotensin II receptor antagonist losartan on appetitive versus aversive learning: a randomized controlled trial. *Biol. Psychiatry* **86**, 397–404 (2019).
69. Huckins, L. M. et al. Analysis of genetically regulated gene expression identifies a prefrontal PTSD gene, *SNRNP35*, specific to military cohorts. *Cell Rep.* **31**, 107716 (2020).
70. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
71. Girgenti, M. J. et al. Transcriptomic organization of the human brain in post-traumatic stress disorder. *Nat. Neurosci.* **24**, 24–33 (2021).
72. Le-Niculescu, H. et al. Towards precision medicine for stress disorders: diagnostic biomarkers and targeted drugs. *Mol. Psychiatry* **25**, 918–938 (2020).
73. Smith, A. K. et al. Epigenome-wide meta-analysis of PTSD across 10 military and civilian cohorts identifies methylation changes in AHRR. *Nat. Commun.* **11**, 5965 (2020).
74. Li, M. et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, eaat7615 (2018).

Department of Veterans Affairs Cooperative Studies Program (no. 575B)

Murray B. Stein[1], Daniel F. Levey[4], Zhongshan Cheng[4], Frank R. Wendt[4], Gita A. Pathak[4], Krishnan Radhakrishnan[9], Mihaela Aslan[9], Hongyu Zhao[9], Renato Polimanti[4], John Concato[9] and Joel Gelernter[4]

A full list of members and their affiliations appears in the Supplementary Information.

VA Million Veteran Program

Murray B. Stein[1], Daniel F. Levey[4], Zhongshan Cheng[4], Frank R. Wendt[4], Kelly Harrington[6,7], Gita A. Pathak[4], Kelly Cho[6], Rachel Quaden[6], Yuk-Lam Anne Ho[6], Daniel Posner[6], Renato Polimanti[4], John Concato[9] and Joel Gelernter[4]

A full list of members and their affiliations appears in the Supplementary Information.

Methods

Subjects. All subjects are enrollees in the MVP[18]. Active users of the Veterans Health Administration healthcare system learn of MVP via an invitational mailing and/or through MVP staff while receiving clinical care, with informed consent and Health Insurance Portability and Accountability Act of 1996 authorization as the only inclusion criteria. As of July 2020, >825,000 veterans have enrolled in the program; for the current analyses, genotype data were available from approximately 375,000 participants. Individuals with an EHR diagnosis of schizophrenia or bipolar disorder were excluded from participation in this study of PTSD. Research involving MVP is approved by the VA Central Institutional Review Board (IRB); the current project was also approved by VA IRBs in Boston, San Diego and West Haven.

PTSD case-control (binary) EHR-derived phenotype. Details on the validation and psychometric properties of this phenotype are reported in our recent publication[23]. In brief, we used manual chart review ($n=500$) as the gold standard. For both the algorithm and chart review, three classifications were possible: probable PTSD, possible PTSD or no PTSD. We used Lasso regression with cross-validation, first to select statistically significant predictors of PTSD from the EHR and then to generate a predicted probability score of being a PTSD case for every participant in the study population. Probability scores ranged from 0 to 1. Comparing the performance of our probabilistic approach (Lasso algorithm) to a rule-based approach (International Classification of Diseases (ICD) algorithm), the former showed modestly higher overall percentage agreement with the latter compared to the ICD algorithm (80 versus 75%), higher sensitivity (0.95 versus 0.84) and higher overall accuracy (area under the curve = 0.95 versus 0.90). For purposes of the case-control binary EHR-derived phenotype used here, we applied a cutoff point of $P=0.7$ to the Lasso results to determine final PTSD case-control status; we also selected a threshold score of 30 on the PCL from the MVP survey to minimize false-negative classifications (for example, due to an absence of PTSD screening information in the EHR). This final algorithm had 0.96 sensitivity, 0.98 specificity, 0.91 positive predictive value and 0.99 negative predictive value for PTSD classification in the transancestral sample as determined by chart review.

PTSD symptom severity (quantitative trait) subphenotypes. The second optional questionnaire, the MVP Lifestyle Survey, includes the PTSD Symptom Checklist (PCL; DSM-IV version)[26], which asks respondents to report the extent to which they had been affected in the previous month by symptoms in response to stressful life experiences. The PCL has 17 items, each scored on a five-point severity scale (1 = Not at all to 5 = Extremely). The re-experiencing (REX) symptom domain is covered by five items (score range 5–25), the avoidance (AVOID) domain by seven items (score range 5–35) and the hyperarousal (HYPER) domain by five items (score range 5–25), yielding an overall severity score (TOTAL) for the 17 items (score range 17–85). After accounting for missing phenotype data, the final sample size for TOTAL was 186,689 in the EUR sample and 25,318 in the AFR sample.

Genotyping, imputation and quality control. Genotyping, imputation and quality control within MVP have previously been described[18]. Briefly, samples were genotyped using a 723,305-SNP Affymetrix Axiom biobank array, customized for MVP. Imputation was performed with minimac3 (ref. [75]) using data from the 1000 Genomes Project. For postimputation quality control, SNPs with imputation INFO scores of <0.3 or MAF < 0.01 were removed from analysis. For the first tranche of data, 22,183 SNPs were selected through LD pruning using PLINK[24,76], and Eigensoft[77] was then used to conduct principal component analysis on 343,286 and 2,504 MVP and1000 Genomes Project samples, respectively[78]. The reference population groups in the 1000 Genomes samples were used to define the groups EUR ($n=241,541$) and AFR ($n=61,796$) used in these analyses. Similar methods were used in the second data tranche, which contained 108,416 new MVP samples and the same 2,504 1000 Genomes Project samples. In the second tranche, 80,694 participants were defined as EUR and 20,584 as AFR. In this manuscript, we report results as the meta-analysis of data from both tranches, either for EUR and AFR separately or as a transancestral meta-analysis.

Association analyses. Genome-wide association studies analysis was carried out by either logistic (for the two binary traits) or linear (for the quantitative traits) regression for each ancestry group and tranche using PLINK 2.0 (ref. [24]) on dosage data, covarying for age, sex and the first ten PCs. Meta-analysis was performed using METAL[25]. We applied a standard genome-wide multiple testing correction ($P<5\times10^{-8}$). No additional multiple testing correction was applied with respect to the number of phenotypes tested, due to their high genetic correlation ($r_g>0.9$). The association results were populated and visualized using Phenogram[79]. Risk loci were enumerated using FUMA[80], and each locus containing more than ten SNPs was fine-mapped using CAVIAR[27] for PCL-Total in the EUR population only, because no significant associations were observed in the AFR population; and using EHR case-control phenotypes for both populations. To understand the biological effect of SNPs associated with PTSD phenotypes, we analyzed the top SNPs (at suggested threshold $P<5\times10^{-6}$) for their unique and overlapping distribution across the five phenotypes. The top SNPs for each phenotype were LD pruned ($r^2=0.2$, kilobases (kb) = 250) to obtain independent signals, and investigated for their role as QTL for protein expression, DNA methylation and splicing (brain tissues) from QTLbase[29].

LDSC and SNP-based heritability. Single-nucleotide polymorphism heritability was calculated using LDSC[81] on the observed scale for continuous phenotypes, and on the liability scale (using prevalence of 10%) for the PTSD case-control definition. Genetic correlation was estimated among PTSD case-control, PCL-Total and all phenotypes from UK Biobank with suitable h^2 accuracy for reliable r_g estimation ($h^2\ z\geq4$). Heritability and genetic correlation analyses were performed using the 1000 Genomes Project European LD reference panel.

Conditional analysis with other psychiatric disorders. Considering the extensive comorbidity between major depression and PTSD[82], we conducted conditional analysis with mtCOJO[30] using genome-wide complex trait analysis software, with the MVP PCL-Total symptom severity summary statistics as the primary analysis and the PGC MDD2 (excluding 23andMe due to data unavailability)[83] summary statistics to condition the analysis for depression. Additional summary statistics for autism spectrum disorder, anorexia nervosa, anxiety (case-control), alcohol dependence, schizophrenia, bipolar disorder and attention deficit hyperactivity disorder were obtained from https://www.med.unc.edu/pgc/results-and-downloads/.

Genomic structural equation modeling. Genomic structural equation modeling (GenomicSEM) was performed in R using the GenomicSEM package[84]. Multivariable linkage disequilibrium matrices were created using the 1000 Genomes Project Phase 3 European reference. Exploratory factor analysis was used to estimate the most appropriate number of latent factors represented by the psychiatric phenotypes and psychopathologies tested, assuming a maximum number of latent factors equal to $n_{traits}-1$. Confirmatory factor analysis was used to calculate factor loadings onto each latent factor(s). Standardized loading values are reported.

PRS analysis. The PRS (Supplementary Fig. 3) were calculated after using *P*-value-informed clumping with an LD cutoff of $r^2=0.05$ within a 500-kb window, excluding the major histocompatibility complex region of the genome because of its complex LD structure. The European samples of the 1000 Genomes Project were used as the LD reference panel. PRS analysis was conducted based on GWAS summary association data using the gtx R package incorporated in PRSice v.1.25 software[81]. For each PRS analysis, we calculated an approximate estimate of the explained variance from a multivariate regression model[85]. For comparison of cross-ancestry PRS (Supplementary Fig. 4) we clumped summary statistics from a recent PTSD GWAS[13], applying an LD cutoff of $r^2=0.3$ within a 500-kb window. These clumped summary statistics were used as a base for calculation of PRS in MVP individuals of EUR and AFR ancestry, independently, using PRSice v.2.0 software[86].

PrediXcan-S methods. To perform transcriptome-wide association analysis, PrediXcan-S (also known as MetaXcan)[41] was used to impute gene expression based on GWAS summary statistics of PCL-Total with the reference gene expression data of 48 tissues from GTEx Release v.7. Gene expression association with PTSD PCL-Total was performed individually for each tissue (13 of which are brain tissues).

Colocalization analysis. Colocalization analysis was performed using the coloc R package[42] for genes that were significant according to the transcriptome-wide association study results of brain tissues with gene expression data from GTEx Release v.8. The coloc.abf function was used to test for shared causal loci under four alternative hypotheses. Loci with posterior probability >90% were considered as strong evidence for the H_4 hypothesis—that is, both traits are associated and share a single causal variant.

Drug repositioning analysis. CMap (https://clue.io/cmap) provides expression similarity scores for a specific expression profile with other drug-induced transcriptional profiles, including consensus transcriptional signatures of 83 drug classes—that is, transcriptional profiles induced by 2,837 drugs grouped into 83 drug classes. Expression similarity is evaluated by means of scores that vary from −100 to 100, with −100 being the most extreme opposite expression profile and 100 the most extreme similar expression profile.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The GWAS summary statistics generated and/or analyzed during the current study will be made available via dbGAP; the dbGaP accession assigned to the MVP is phs001672.v1.p. The website is https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v1.p1.

References

75. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
76. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
77. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
78. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
79. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).
80. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
81. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
82. Koenen, K. C. et al. Common genetic liability to major depression and posttraumatic stress disorder in men. *J. Affect. Disord.* **105**, 109–115 (2008).
83. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
84. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
85. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
86. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).

Acknowledgements

This research is based on data from the MVP, Office of Research and Development, Veterans Health Administration and was supported by funding from the VA Cooperative Studies Program (CSP, no. CSP575B) and the Veterans Affairs Office of Research and Development MVP (grant nos. MVP000 and VA Merit MVP025). The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government. We thank the veterans who participated in this study, and the members of the VA CSP and MVP study teams, without whom this work would not have been possible.

Author contributions

M.B.S. and J.G. had primary responsibility for design of the study. M.B.S., J.G., J.C., K.R. and M.A. supervised the study and managed and organized the group. D.F.L., Z.C., F.R.W., G.A.P. and R.P. contributed to genetic and bioinformatic analyses. K.H., K.C., R.Q., Y.-L.A.H., K.R., M.A. and D.P. contributed to phenotyping and phenomic analyses. The initial manuscript was drafted by M.B.S., D.F.L., R.P. and J.G. Manuscript contributions and interpretation of results were provided by M.B.S., D.F.L., Z.C., F.R.W., G.A.P., K.H., M.J.G., D.P., R.S.D., H.Z., R.P., J.C. and J.G. The remaining authors contributed to other organizational or data-processing components of the study. All authors saw, had the opportunity to comment on and approved the final draft.

Competing interests

M.B.S. has in the past 3 years been a consultant for Actelion, Acadia Pharmaceuticals, Aptinyx, Bionomics, BioXcel Therapeutics, Clexio, EmpowerPharm, Epivario, GW Pharmaceuticals, Janssen, Jazz Pharmaceuticals, Roche/Genentech and Oxeia Biopharmaceuticals. M.B.S. has stock options in Oxeia Biopharmaceuticals and Epivario. J.G. is named as coinventor on PCT patent application no. 15/878,640, entitled 'Genotype-guided dosing of opioid agonists', filed 24 January 2018. None of the other authors declare any competing interests.

Additional information

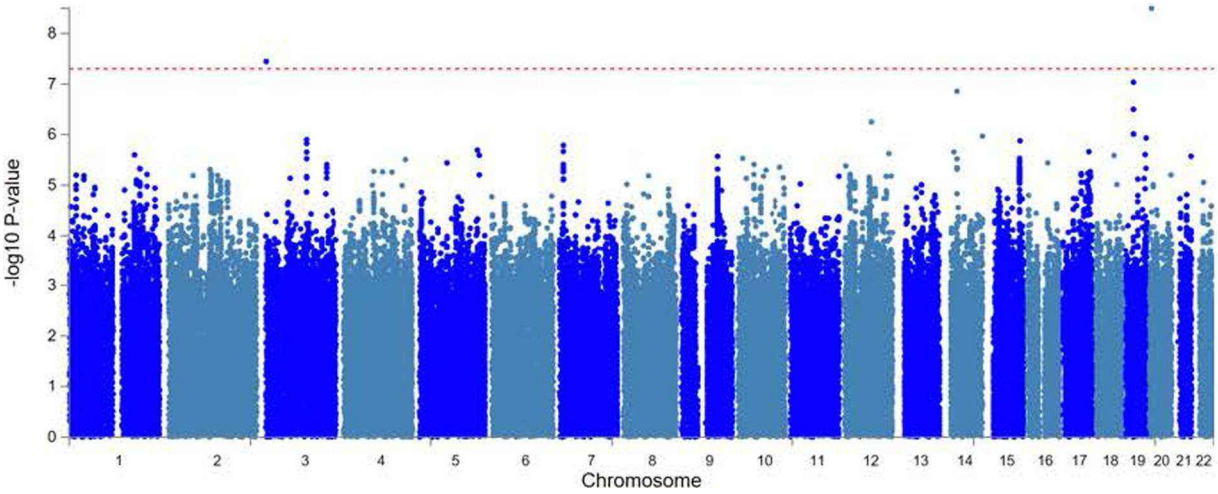
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00767-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00767-x.

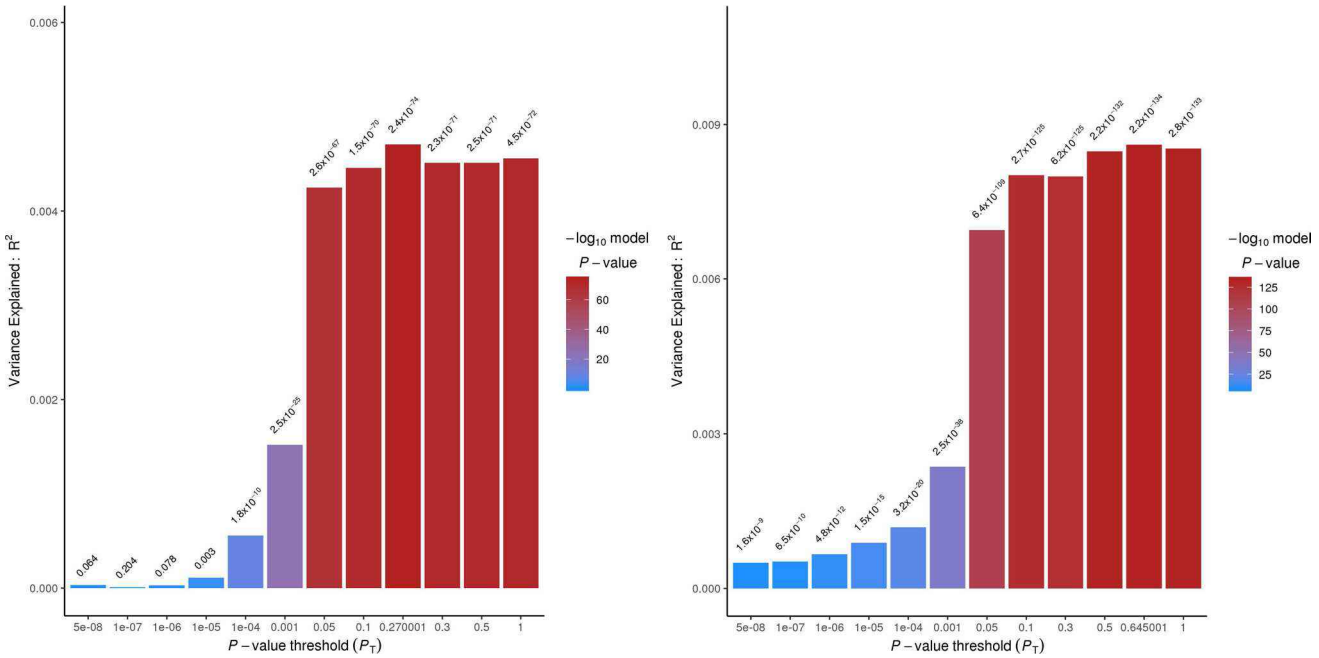
Correspondence and requests for materials should be addressed to M.B.S. or J.G.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

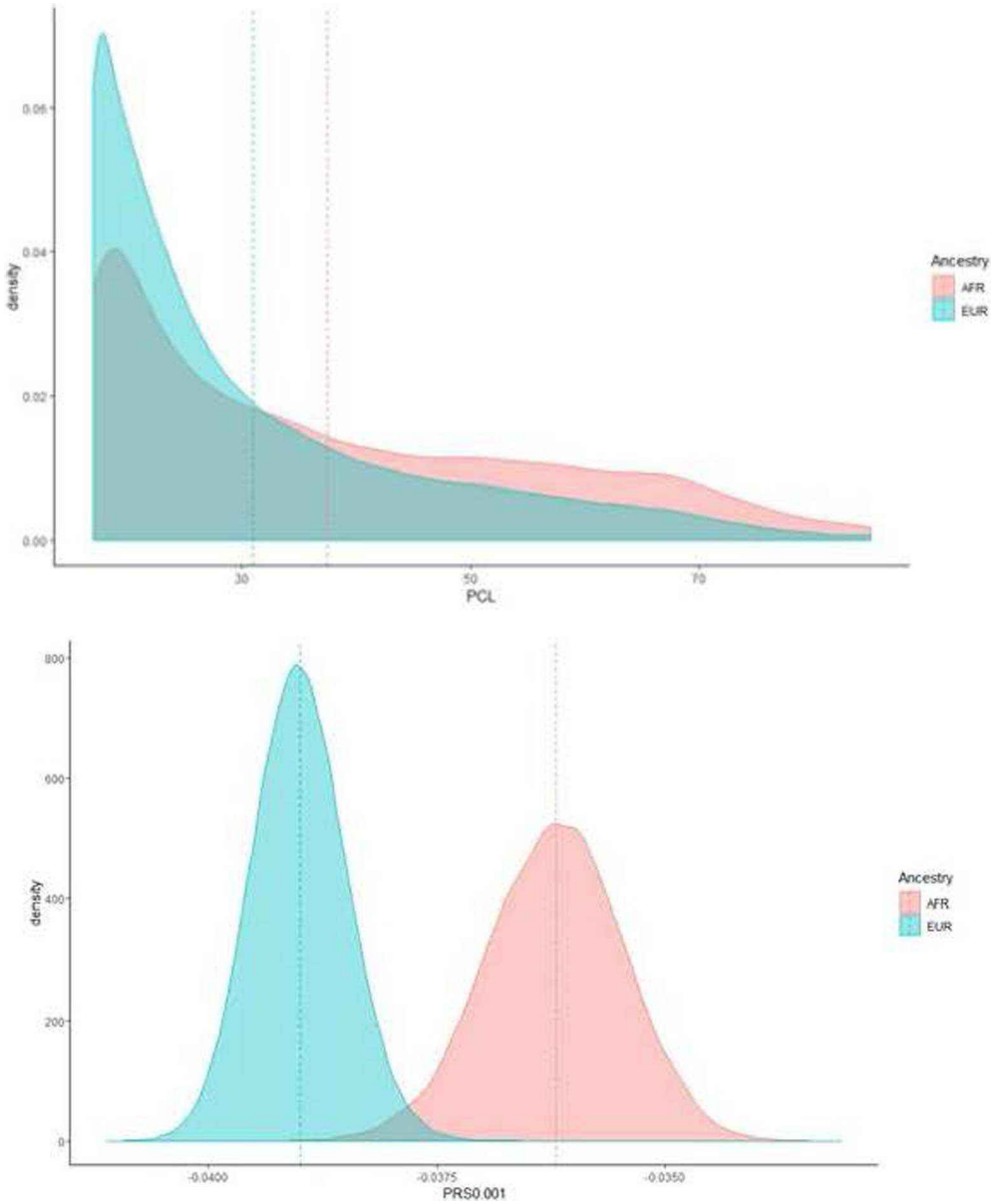
Reprints and permissions information is available at www.nature.com/reprints.



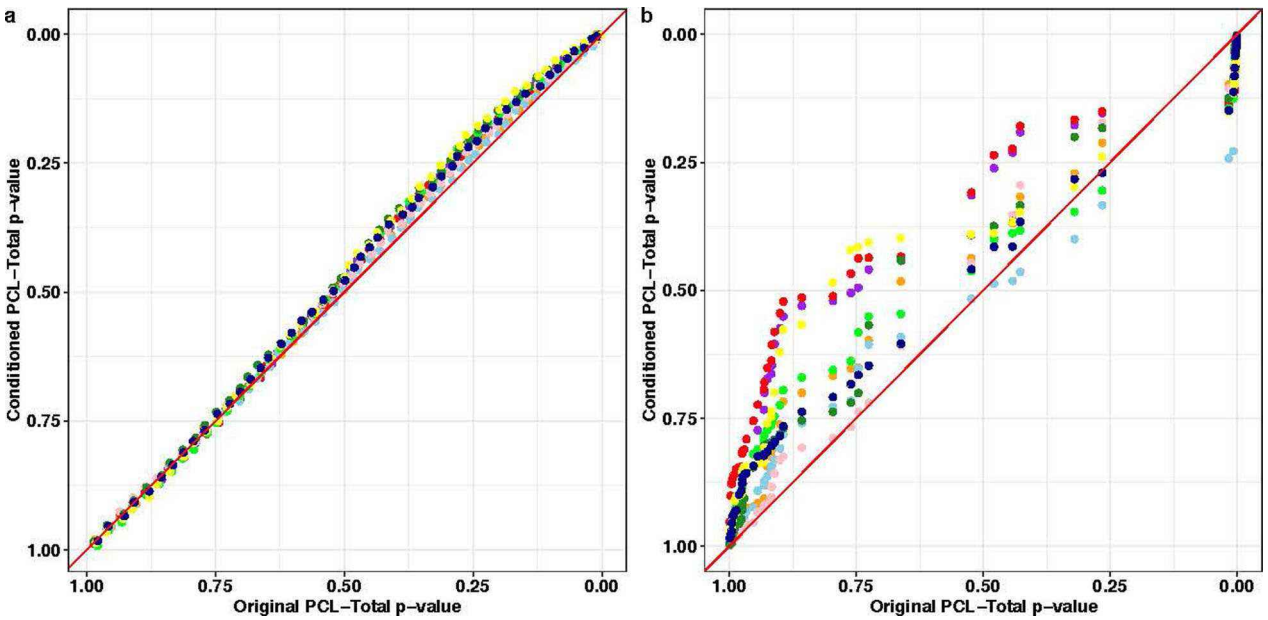
Extended Data Fig. 1 | Manhattan plot of MVP AFR case-control GWAS. Horizontal red line indicates $P < 5 \times 10^{-8}$. *P*-values are uncorrected. Results are based on logistic regression.



Extended Data Fig. 2 | Polygenic risk scores in MVP and PGC-PTSD. Polygenic risk score (PRS) from MVP EUR case-control (left) and EUR PCL-total (right) applied to PGC-PTSD[13] case-control phenotype with varying *P*-value thresholds (PT) on the *x*-axis and explained variance (R^2) on the *y*-axis. The approximate estimate of the explained variance was calculated using a multivariate regression model. *P* values reported are two sided, and Bonferroni correction accounting for the number of *P*-value thresholds tested is $P = 2.38 \times 10^{-4}$.



Extended Data Fig. 3 | Symptom and polygenic risk scores in veterans of African and European ancestry. Top shows density plot of PCL-total scores in veterans of AFR (salmon color) and EUR (teal color) ancestry. Bottom shows density plot of PRS scores (at *P*-value threshold 0.001) for MVP PCL AFR (salmon color) and MVP PCL EUR (teal color) derived from PGC PTSD EUR.



Extended Data Fig. 4 | Gene Ontology (GO) term and GTEx tissue enrichment. a, Quantile-quantile plots between Gene Ontology (GO) term enrichment (one-sided test for positive relationship between tissue and genetic association) in original PCL-Total and conditioned PCL-Total (blue, autism spectrum disorder; purple, major depression; dark green, anorexia nervosa; light green, anxiety; pink, schizophrenia; light blue, bipolar disorder; orange, attention deficit hyperactivity disorder; red, all eight disorders simultaneously). **b**, Quantile-quantile relationship between GTEx tissue enrichment (one-sided test for positive relationship between tissue and genetic association) in original PCL-total and conditioned PCL-Total. To avoid over-plotting, enrichment *P*-values were divided into quantiles. Red diagonal lines indicate a one-to-one relationship between original and conditioned PCL-Total gene set and tissue enrichments. Two-sided tests were used to compare enrichment results.

nature research

Corresponding author(s): Stein MB and Gelernter J

Last updated by author(s): Dec 5, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Details on PTSD phenotyping in the MVP cohort have been described previously, see Harrington et al. 2019 (PMID: 31009556).
Data analysis	Minimac3, PLINK version 2.0, flashpca, RVTEST, PrediXcan-S, LD score regression (LDSC), LDHub v2.0, PRSice v1.25 and v2.0, GTEx Analysis Release V7 and V8, eQTL, 1,000 Genomes Project Reference data, R packages (GenABEL, GenomicSEM, GLMNET [for Lasso analyses], psych), mtCOJO using GCTA sofware, Eigensoft, CAVIAR, FUMA v1.34, METAL, CMap (https://clue.io/cmap), SAS version 9.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The GWAS summary statistics generated during and/or analyzed during the current study will be made available via dbGAP; the dbGaP accession assigned to the Million Veteran Program is phs001672.v1.p. The website is: https:// www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v1.p1

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National*

Research sample	*Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Human research participants
☐	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*

Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.