SUPPLEMENTARY FILE

Title: PTSD blood transcriptome mega-analysis:
Shared inflammatory pathways across biological sex and modes of trauma

Authors: Michael S. Breen,[1,2*] Daniel S. Tylee,[3] Adam X. Maihofer[4], Thomas C. Neylan[5,6], Divya Mehta[7], Elisabeth Binder[8,9], Sharon D. Chandler[4], Jonathan L. Hess[3], William S. Kremen[4,10], Victoria B. Risbrough[4,10], Christopher H. Woelk[11], Dewleen G. Baker[4,10], Caroline M. Nievergelt[4,10], Ming T. Tsuang[4,10], Joseph D. Buxbaum[1,2] and Stephen J. Glatt[3]

Affiliations: [1]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [2]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [3]Psychiatric Genetic Epidemiology & Neurobiology Laboratory (PsychGENe Lab); Departments of Psychiatry and Behavioral Sciences & Neuroscience and Physiology; SUNY Upstate Medical University, Syracuse, NY, USA; [4]Department of Psychiatry, University of California San Diego, California, USA; [5]Department of Psychiatry, University of California San Francisco, California, USA; [6]San Francisco Veterans Affairs Medical Center; [7]School of Psychology and Counseling, Faculty of Health, Queensland University of Technology, Kelvin Grove, Queensland, Australia; [8]Department of Translational Research in Psychiatry, Max-Planck Institute of Psychiatry, Munich, Germany; [9]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA; [10]Veterans Affairs Center of Excellence for Stress and Mental Health, San Diego, California, USA; [11]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, UK.

* To whom correspondence should be addressed:

Michael S. Breen
Department of Psychiatry
Icahn School of Medicine at Mount Sinai
New York, NY, 10029, USA
Tel: +1 (212) 241-0242
Email: michael.breen@mssm.edu

This Supplementary File contains additional details on methods and results.

METHODS

Gene expression data processing and quality control

All statistical analyses were conducted in the statistical package *R*. Data from each study were processed, normalized and quality treated independently. All microarray data were quantile-normalized and $\log_2$ transformed. Affymetrix arrays underwent robust multi-array average normalization (Carvalho *et al*, 2006) with additional GC-correction when possible (Gautier *et al*, 2004; Carvalho *et al*, 2010; Wu and Gentry, 2016) and Illumina and CodeLink arrays were treated as described previously (Neylan *et al*, 2011). When multiple microarray probes mapped to the same HGNC symbol, the expression of the probe with the highest average value across all samples was used for subsequent analysis. RNA-sequencing count data were treated as described previously (Breen *et al*, 2015) using the VOOM normalization (Ritchie *et al*, 2015), a variance-stabilization transformation method resulting in a normally distributed data matrix. Normalized data were inspected for outlying samples using unsupervised hierarchical clustering of samples (based on Pearson's coefficient and average distance metric) and principal component analysis to identify potential outliers outside two standard deviations from these grand averages. Combat batch correction (Leek *et al*, 2015) was applied to remove systematic sources of variability other than case/control status, such as technical, clinical, or demographic factors both within each study (as necessary), and then across all studies using common gene symbols, forming the bases for subsequent mega-analytic case-control groups.

Peripheral blood cell type estimates

The frequencies of circulating immune cells were estimated for each individual in each study using Cibersort cell type de-convolution (https://cibersort.stanford.edu/) (Newman *et al*, 2015). Cibersort relies on known cell subset specific marker genes and applies linear support vector regression, a machine learning approach highly robust compared to other methods with respect to noise, unknown mixture content and closely related cell types. As input, we used the LM22 signature matrix to distinguish seven main leukocytes subtypes: B cells, cytotoxic T cells ($CD8^+$), helper- and regulatory T cells ($CD4^+$), natural killer (NK) cells ($CD56^+$), monocytes ($CD14^+$), eosinophils and neutrophils. The LM22 matrix can be further divided into 14 less frequent immune cell subsets, which we pooled and defined as 'other'. The resulting estimates were tested for normality using Kolmogorov-Smirnov test and a two group Wilcoxon Signed Rank tests with *post hoc* Tukey correction was used compare means between PTSD cases and controls for each study.

Functional annotation and protein interaction networks

All significant DGE signatures and gene modules were subjected to functional annotation. First, the ToppFunn module of ToppGene Suite software (Chen *et al*, 2015) was used to assess enrichment of Gene Ontology (GO) terms specific to biological processes and molecular factors using a one-tailed hyper geometric distribution with family-wise false discovery rate (FDR) at 5%. GO semantic similarity analysis was used to assess shared/unique gene content amongst GO terms using the GoSemSim semantic similarity R package (Yu *et al*, 2015), and default semantic contribution factors ('is_a' relationship: 0.8 and 'part_of' relationship: 0.5). This analysis results in a symmetric matrix in which each value represents a score for similarity between GO term pairs. Then, we undertook hierarchical clustering based on semantic similarity matrix to group together all GO terms with common GO 'parent'. Second, gene modules were tested for over-representation of PTSD genome-wide association study (GWAS) signatures obtained from the DisGenNet database (Pinero *et al*, 2015), retrieved using the disease-term query 'PTSD'. Third, DGE signatures were used to build direct protein-protein interaction (PPI) networks, which can reveal key genes/transcription factors mediating the regulation of multiple target genes. PPIs were obtained from the STRING database (Franceschini *et al*, 2012) with a signature query of DGE lists from the mega-analytic case-control comparisons. STRING implements a scoring scheme to report the confidence level for each direct PPI (low confidence: <0.4; medium: 0.4–0.7; high: >0.7). We used a combined STRING score of >0.4. We further used STRING to test whether the number of observed PPIs were significantly more than expected by chance using a nontrivial random background model (that is, null model). For visualization, the STRING network was imported into CytoScape (Shannon *et al*, 2003).

Cross-disorder overlap analyses

To compliment our study, we gathered gene-level statistics from other recent blood transcriptome mega-analyses of schizophrenia (Hess *et al*., 2016) and autism spectrum disorder (Tylee *et al*., 2016), as well as emerging, unpublished data on bipolar disorder (Hess *et al*., in prep.) We also gathered dysregulated genes from a recent large-scale blood transcriptome investigation of major depressive disorder (Jansen *et al*., 2016). For each disorder, disease-related genes with a *P*-value <0.05 were used to yield a sufficient number of genes to perform a gene ontology over-representation analysis (as above). Subsequently, to determine specificity to PTSD,

we performed a series of cross-disorder overlaps at both the individual gene and gene-ontology level using a two-tailed Fisher's exact test.

Construction of PTSD blood-based diagnostic classifiers

BRB-Array Tools supervised classification methods (Simon *et al*, 2007) were used to construct gene expression classifiers. Three models were specified to distinguish PTSD cases from controls relative to: (1) men exposed to combat trauma (2) men exposed to IP traumas, and (3) women exposed to IP traumas. Each model consisted of three steps. First, to ensure a fair comparison, all genes in the training data with $P < 0.05$ were subjected to classifier construction, respective for each mega-analytic case-control group. This heuristic rule of thumb approach was used to cast a wide net to catch all potentially informative genes, while false-positives would be pared off by subsequent optimization and cross-validation steps. Second, classifiers composed of different numbers of genes were constructed by recursive feature elimination (RFE). RFE provided feature selection, model fitting and performance evaluation via identifying the optimal number of features with maximum predictive accuracy. RFE selected the top 100 differentially expressed genes and evaluated classification accuracies within the training data (70% of data) using a two-layer leave one-tenth out cross-validation approach, prior-to predicting class labels on completely withheld test data (30% of data). This process iteratively tested classification accuracies by removing the five least predictive genes. Third, the ability for RFE to predict group outcome was assessed by support vector machines (SVM) and compared to four different multivariate classification methods (*i.e.* diagonal linear discriminant analysis (DLDA), nearest centroid (NC), first-nearest neighbors (1NN), three-nearest neighbors (3NN)). For each of the three models, classification accuracies are reported for both the training data and the completely withheld test data as area under the receiver operating curve (AUC).

RESULTS

Stratified gene co-expression module preservation analyses

Between trauma-type comparisons: Initially, weighted gene co-expression network analysis (Langfelder and Horvath, 2008) (WGCNA) was used to assess whether exposure to different traumatic events may influence gene co-regulatory patterns as being disrupted or created when exposed to combat traumas relative to IP traumas, and *vice versa*. These between trauma-type comparisons were first assessed within trauma-exposed control individuals and then separately within PTSD cases using a

permutation-based preservation statistic ($Z_{summary}$). In trauma-exposed control individuals, low preservation statistics ($Z_{summary}$ <2) were observed for one module when comparing those with a history of combat traumas relative to those with a history of IP traumas, while five modules with low preservation statistics were observed when we tested the reverse relationship (**Supplementary Figure 4A**). In contrast, moderate-to-high preservation statistics were observed when comparing PTSD individuals with a history of combat traumas relative to those with a history IP traumas, and *vice versa* (**Supplementary Figure 4B**), suggesting that the molecular response to different traumatic events are possibly more homogenous in trauma survivors with PTSD compared to non-PTSD controls (P=0.01) (**Supplementary Figure 4C**).

REFERENCES

1. Breen MS, Maihofer AX, Glatt SJ, Tylee DS, Chandler SD, Tsuang MT et al (2015). Gene networks specific for innate immunity define post-traumatic stress disorder. Molecular Psychiatry 20: 1538-1545.
2. Carvalho B, Bengtsson H, Speed T, Irizarry, R (2006). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 8: 485-499.
3. Carvalho B, Irizarry R (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics 26: 2363-2367.
4. Chen J, Bardes E, Aronow B, Jegga A (2015). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research 37: W305-W311.
5. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A et al (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research 41: D808-D815.
6. Gautier L, Cope L, Bolstad B, Irizarry R (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307-315.
7. Hess J, Tylee DS, Barve R, de Jong S, Ophoff RA, Kumarasinghe N et al (2016). Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. Schizophrenia Research 176: 114-124.
8. Jansen R, Penninx BW, Madar V, Xia K, Milaneschi Y, Hottenga JJ et al (2016). Gene expression in major depressive disorder. Molecular Psychiatry 21(3): 339–47.
9. Langfelder P, Horvath S (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.
10. Leek J, Johnson W, Parker H, Jaffe A, Storey J (2015). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882-883.
11. Newman A, Long Liu A, Green MR, Gentles AJ, Feng W, Xu Y et al (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 12: 453-457.

12. Neylan T Sun B, Rempel H, Ross J, Lenoci M, O'Donovan A et al (2011). Suppressed monocyte gene expression profile in men versus women with PTSD. Brain, Behavior, and Immunity 25: 524-53.
13. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13: 2498-2504.
14. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y (2007). Analysis of gene expression data using BRB-Array tools. Cancer Inform 2: 11-17.
15. Tylee DS, Jess JL, Quinn TP, Barve R, Huang H, Zhang-James Y et al (2016). Blood Transcriptomic Comparison of Individuals With and Without Autism Spectrum Disorder: A Combined-Samples Mega-Analysis. Am J Med Genet Part B 9999: 1–21.
16. Wu J, Gentry RI (2016). gcrma: Background Adjustment Using Sequence Information. R package version 2.46.0.
17. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S (2015). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26: 976-978.