Priority Communication

Biological Psychiatry

Genome-wide Association Study of Dimensional Psychopathology Using Electronic Health Records

Thomas H. McCoy Jr., Victor M. Castro, Kamber L. Hart, Amelia M. Pellegrini, Sheng Yu, Tianxi Cai, and Roy H. Perlis

ABSTRACT

BACKGROUND: Genetic studies of neuropsychiatric disease strongly suggest an overlap in liability. There are growing efforts to characterize these diseases dimensionally rather than categorically, but the extent to which such dimensional models correspond to biology is unknown.

METHODS: We applied a newly developed natural language processing method to extract five symptom dimensions based on the National Institute of Mental Health Research Domain Criteria definitions from narrative hospital discharge notes in a large biobank. We conducted a genome-wide association study to examine whether common variants were associated with each of these dimensions as quantitative traits.

RESULTS: Among 4687 individuals, loci in three of five domains exceeded a genome-wide threshold for statistical significance. These included a locus spanning the neocortical development genes *RFPL3* and *RFPL3S* for arousal ($p = 2.29 \times 10^{-8}$) and one spanning the *FPR3* gene for cognition ($p = 3.22 \times 10^{-8}$).

CONCLUSIONS: Natural language processing identifies dimensional phenotypes that may facilitate the discovery of common genetic variation that is relevant to psychopathology.

Keywords: Arousal, Genetic, Genomic, Social, Transdiagnostic, Valence

https://doi.org/10.1016/j.biopsych.2017.12.004

Family studies of psychiatric illnesses demonstrated decades ago the overlap in risk for these disorders, a finding that has now been confirmed by genome-wide association studies (1–3). Such overlap highlights the limitations of a nosologic system focused on categories of symptoms rather than dimensions. For this reason, recent initiatives emphasize the utility of identifying symptom domains that may better correspond to underlying neurobiology (4,5).

The rise of biobanks embedded in health care systems or national registries provides an opportunity to investigate the impact of genomic variation in a less biased fashion than traditional disease case-control designs. However, such biobanks typically capture primarily coded clinical data, i.e., categorical diagnoses. We have recently developed multiple methods to examine narrative clinical notes to extract symptom dimensions as a means of augmenting these coded data (6,7).

We hypothesized that symptom dimensions based on expert-curated terms capturing National Institute of Mental Health Research Domain Criteria (RDoC) domains would be associated with common genomic variation and could thereby implicate novel sets of genes related to psychopathology. As proof of concept, we therefore applied a newly described natural language processing (NLP) method for extracting dimensional phenotypes to hospital discharge summaries drawn from the genomic biobank of an academic medical center (7) and used standard genome-wide association studies to investigate these novel phenotypes as quantitative traits.

METHODS AND MATERIALS

Overview and Data Set Generation

We drew on three waves of participants in the Partners Biobank from the Brigham and Women's Hospital network and the Massachusetts General Hospital network, representing approximately the first 15,000 individuals genotyped as part of the Partners HealthCare Biobank initiative (8). Narrative discharge summaries were extracted from the longitudinal electronic health record of the Massachusetts General Hospital. We included any individuals 18 years of age or older who had at least one hospitalization between 2010 and 2015.

A datamart containing all clinical data was generated with i2b2 server software (version 1.6; i2b2, Boston, MA), a computational framework for managing human health data (9–11). The Partners HealthCare System Institutional Review Board approved both the study protocol and the release of biobank data, which were collected after acquiring written informed consent from participants and explicitly allowed identifiable data to be shared with qualified investigators.

ISSN: 0006-3223

Study Design and Analysis

Primary analyses used a cohort design with all patients admitted for any reason during the time period noted above. Discharge documentation was used to estimate dimensional psychopathology scores for one encounter per individual; when an individual was hospitalized on multiple occasions during the study period, a single hospitalization was selected at random to minimize bias resulting from other means of ascertainment. The derivation of dimensional psychopathology has been described elsewhere (7); in brief, it began with a set of seed terms for each of the five National Institute of Mental Health RDoC definitions drawn from National Institute of Mental Health workgroup statements, then expanded these term lists to include synonyms (12). This second expansion step is important because it reduces potential bias introduced by a given specialty or set of providers who may use specific terminology to characterize symptoms, yielding a broader set of terms that should better generalize across providers and hospitals. Each note is assigned a score corresponding to a simple count of term appearance. We have developed simple code to facilitate dimension extraction in other data sets (7).

Genotyping and Quality Control

DNA was extracted from buffy coat, and genotyping was done using three versions of the Illumina Multi-Ethnic Global (MEG) array (Illumina, Inc., San Diego, CA) (MEGA, n = 4927; MEGA EX, n = 5353; and MEG, n = 4784; mappable variants available for each were 1,411,334, 1,710,339, and 1,747,639, respectively). These common variant arrays all incorporate content from the 1000 Genomes Project Phase 3. Single nucleotide polymorphism (SNP) coordinates were remapped based on the TopGenomicSeq provided by Illumina; all reference SNP cluster IDs correspond to build 142 of the Single Nucleotide Polymorphism Database. To determine the forward strand of the SNP, we aligned both SNP sequences (alleles A and B) to hg19 using the BLAST-like alignment tool (BLAT) with default parameters set by the University of California Santa Cruz Genome Browser (13).

Each cohort was cleaned, imputed, and analyzed separately to avoid batch effects. In each batch we included subjects with genotyping call rates exceeding 99%; no related individuals based on identity by descent were included (14). From these individuals, any genotyped SNP with a call rate of at least 95% and a Hardy-Weinberg equilibrium p value $<1 \times 10^{-6}$ was included. Imputation used the Michigan Imputation Server implementing Minimac3 (15–17). Imputation used all population subsets from 1000 Genomes Project Phase 3 version 5 as reference panel; haplotype phasing was performed using SHAPEIT (18).

For each batch, we applied principal components analysis of a linkage-disequilibrium-pruned set of genotyped SNPs to characterize population structure, based on EIGENSTRAT as implemented in PLINK software version 1.9 (19). We then plotted these components with superimposition of HapMap samples to confirm location of Northern European individuals. The present analysis included only individuals of Northern European genomic ancestry to minimize the risk for confounding by ancestry (i.e., population stratification) and because the power to detect association in other ancestry groups would be limited (20–22).

Analysis

We examined single-locus associations in each batch, then combined in inverse-variance-weighted fixed effects meta-analysis. In all analyses, only biallelic SNPs with minor allele frequencies of at least 1% in all batches were retained. Tests for association used linear regression assuming an additive allelic effect and examined each of the five dimensional measures as a quantitative trait, with adjustment for the first 10 principal components a priori. (In previous work, analyses incorporating five or 20 components did not yield meaningfully different results.) Association results are presented in terms of independent loci after pruning using the clump command in PLINK, with a 250-kb window and r^2 = .2. Locus plots were generated using LocusZoom software (19,23).

Reported p values are not adjusted for lambda or linkage disequilibrium scores; in previous work, adjustment for lambda-1000 or linkage disequilibrium score regression intercept did not meaningfully change relative results. Lambdas ranged from 0.998 to 1.003 (24).

RESULTS

We examined 4687 individuals of Northern European ancestry across the three batches (wave 1, 1589; wave 2, 1547; wave 3, 1551), with meta-analysis of 893,900 SNPs with minor allele frequency of 0.01 or greater. The cohorts included 2363 females (50.4%), and the mean age was 64.3 years (SD, 14.9 years). Figure 1 shows Manhattan plots for each of the five dimensional phenotypes (Q-Q plots are shown in Supplemental Figure S1).

For each of the dimensions, the 10 independent loci with strongest evidence of association are described in Table 1. Overall, one locus was associated with arousal, two with social, and one with cognition at a standard genome-wide significance threshold ($p < 5 \times 10^{-8}$); these four regions are depicted in Figure 2. Notably, for arousal, the associated locus spans *RFPL3* and *RFPL3S*; this family of proteins has been suggested to be important in primate neocortical evolution (25). For cognition, the associated locus spans *FPR3*, a chemoattractant (15623572) that has been suggested to be relevant in immune response in Alzheimer's disease (26).

DISCUSSION

In this analysis of 4687 individuals drawn from a biobank spanning two academic medical centers, we identified four loci associated with dimensional psychopathology at a standard genome-wide threshold based on natural language processing of narrative hospital discharge notes. Two of these span genes are associated with neurodevelopment (*RFPL3*) or neurodegeneration (*PFR3*). While both are known to be brain expressed, neither has previously been strongly associated with neuropsychiatric disease, suggesting the potential utility of the approach we describe in understanding brain function in a manner that is unbiased by traditional nosology.

While not achieving a genome-wide threshold for significance, we also note the observed association between the calcium channel subunit *CACNA2D3* and positive valence.