Multilocus genetic association analyses of suicidal behavior and schizophrenia

Ph.D. thesis

Attila József Pulay, M.D.

Semmelweis University

Doctoral School of Mental Health Sciences

Supervisor:	János Réthelyi, M.D., Ph.D.
Official reviewers:	Péter Antal, Ph.D.
	Gabriella Juhász, M.D., Ph.D.
Head of the Final Examination Commitee:	
	Dániel Bereczki, M.D., Ph.D.
Members of the Final Examination Commitee:	
	Csaba Barta, M.D., Ph.D.
	György Szekeres, M.D., Ph.D.

Budapest

2016

Introduction

It is an age-old observation that certain personality traits or mental problems occur more frequently among blood relatives. Family and twin studies of psychiatric disorders have indicated polygenic inheritance; in some cases (e.g. schizophrenia, autism, bipolar disorder) with high heritability (H^2). The high level of heritability sparked interest in uncovering underlying genetic architecture.

However, psychiatric disorders are diagnosed based on behavioral signs and symptoms. They are also complex phenotypes with quantitative traits with conventional, (i.e. arbitrary) thresholds between physiologic and pathologic functions. Even the most elementary symptoms have such complex neurobiology that it is extremely difficult to accurately predict the effect of genes on behavior. In addition to the circuitous route between phenotype and genotype, the weak effect size (OR<1.2) of single nucleotide polymorphisms (SNP) and variations (SNV) the high genetic heterogeneity and pleitropy stemming from polygenic inheritance also complicates the identification of genetic factors. In this thesis, we analyzed the genetic association of two psychiatric phenotypes with great public health significance; namely, suicidal behavior, and schizophrenia, by using multilocus statistical approaches to fit the polygenic inheritance model.

The Genomic Background of Schizophrenia

Schizophrenia is a complex neuropsychiatric disorder usually starting during young adulthood and characterized by recurring psychotic episodes causing marked dysfunction. The prevalence of schizophrenia is estimated between 0.5-1.2% worldwide, with high heritability (H^2=0.7-0.8), suggesting a predominantly organic etiology independent from culture.

State-of-the-art, genome-wide studies with sufficient sample size identified several variants associated with various genes and pathways depending on their type. The common SNP markers of schizophrenia were found on the non-exonic regions of genes implicated in synaptic plasticity, neurodevelopment, and/or have high therapeutic significance. The rare single nucleotide variation (SNV) and insertion-deletion variants (indels) are overrepresented among genes involved with synaptic plasticity and cognitive function. Association studies with multi-locus statistic approaches indicated association of genes and pathways that play roles in axonal growth, neuronal migration, and immune system function, in addition to synaptic plasticity and neurodevelopment. Polygenic risk-prediction based analyses yielded considerable proportions of polygenic risk shared between schizophrenia, bipolar disorder, major depression, and ADHD.

The Genomic Background of Suicidal Behavior

According to the WHO 2000-2012 Global Health Estimates report, suicide is the second leading cause of death among 15-29 year olds and listed in the top 20 causes of death worldwide. It has a strong relationship with psychiatric disorders; especially mood disorders, that are robust, independent predictors of suicidal ideation and attempts. In addition, suicide is linked so tightly to executive dysfunction that the latter is considered a candidate endophenotype of suicidal behavior.

Suicidal ideations and plans are much more common than actual attempts (lifetime prevalence of suicidal ideation: 9.2%, and suicidal attempts: 2.7%), with low to moderate heritability estimated in genetic epidemiological studies (H^2=17-48%). Genome-wide association studies on suicidal behavior did not find any statistically significant, replicated SNP associations. SNP markers with suggestive associations were identified in genes of the BDNF-CREB-MAPK pathway, the related microRNA (miRNA) genes and in genes involved in synaptic plasticity.

Objectives

Our studies shared the same general objective; namely, to map the genetic architecture of psychiatric disorders with multilocus association tests capable of assessing the joint effect of the analyzed loci as opposed to the standard, SNP-based approaches. The aims of our studies included:

- A cross-disorder, region-based analysis of potentially common candidate genes of suicidal behavior co-occurring with major depressive disorder and bipolar disorder.
- Assessing the shared polygenic risk in suicidal behavior co-occurring with major depressive disorder and bipolar disorder by using PRS scores derived from candidate genes of suicidal behavior and microRNA genes expressed in the prefrontal circuits that are essential in executive functions.
- Hypothesis-free exploration of the genetic architecture of DSM-IV schizophrenia using multilocus association methods.
- Assessing the replication probabilities of various multilocus association methods by using samples of DSM-IV diagnosed schizophrenia with disparate study designs.

Methods

Samples, Phenotypes

Samples from our first study were downloaded from the database of Genotypes and Phenotypes (dbGaP) by the National Center for Biotechnology Information (NCBI) and the Genetic Association and Information Network (GAIN). We analyzed European-American samples of the GAIN Whole Genome Association Study of Bipolar Disorder version 3 (accession id: : phs000017.v3.p1), and samples with Northern-European ancestry from the GAIN Major Depression: Stage 1 Genomewide Association in Population-Based Samples GWAS studies (accession id: phs000020.v2.p1). Analyses were restricted to the cases with or without suicidal behavior in both datasets; samples with schizoaffective disorder, missing phenotypes, healthy controls, samples with high cryptic relatedness, and population outliers were excluded. Based on the available phenotypic information, suicidal behavior was defined in the MDD sample as clear suicidal intent or severe suicidal ideation or attempts (N_{MDD}=1753, N_{suic}=245)), whereas one or more suicidal attempts with ambivalent or clear intent in the bipolar sample (N_{BIP}=999, N_{suic}=358).

The discovery sample of our second study consisted of the DSM-IV schizophrenia case-trio samples of the SCHIZOBANK Project (N=16 trios); most of them (N=12 trios) included patients treated for DSM-IV schizophrenia at the Department of Psychiatry and Psychotherapy of Semmelweis University. The replication dataset of the second study (N=5337) was drawn by merging two case-control GWAS samples of

DSM-IV schizophrenia accessed from the NCBI dbGaP database: The European-American samples of the GAIN Genome-Wide Association Study of Schizophrenia (GAIN SCZ, accession id: phs000021.v3.p2, N=2787), and the Molecular Genetics of Schizophrenia (MGS SCZ, accession id: phs000167.v1.p1, N=2935). During the data cleaning steps, samples not passing QC filters, those with missing phenotypes, with high cryptic relatedness, or population outliers were excluded. Diagnoses of schizophrenia were made according to the DSM-IV criteria in bot datasets, by using clinical diagnoses and structured interview (SCID-I) in the SCHIZOBANK data, as well as the agreement of the most likely diagnosis of at least two experienced psychiatrists in the GAIN and MGS SCZ samples.

Genotyping, Exome Sequencing, Imputation

The genotyping of the GWAS datasets included in this thesis was performed on DNA samples isolated from blood with a Perlegen 600K DNA chip in the MDD sample, and an Affymetrix Genome-Wide Human SNP Array 6.0 DNA chip in the bipolar, GAIN and MGS SCZ samples. Genotyping was followed by standard GWAS quality-control (QC) filtering procedure: sample QC filters included low SNP call-rate (MDD<85%, BIP<98,5%, SCZ < 98%), high cryptic relatedness/duplicated samples, extreme heterozygosity, and differences between genotypic and recorded sex. SNP QC filters included low genotype call-rate (<95%), monomorphic or rare markers (MAF<0.01), significant deviations from Hardy-Weinberg equilibrium (p < 10E-06), and markers with significant batch-effect. In the bipolar sample 729,087

SNPs, in the MDD data 437,114 SNPs, and in the aggregated MGS+GAIN SCZ samples 716,923 SNPs passed the QC filters.

We performed whole-exome sequencing on the DNA samples of the SCHIZOBANK trios by using an Illumina HiScanSQ Next-Generation sequencing platform and NimbleGen SeqCap EZ Human Exome Library v3.0 using 100 bp long, paired-end reads. Reads were aligned with a Burrows-Wheeler Aligner algorithm according to the hg19 reference genome. During the QC steps, genotypes with Mendelian errors, low coverage or quality (depth<4, GQ<20), non-biallelic, singleton or low quality variants (SQ<20, MQ<40), and variants with high missingness (mean+2SD) were filtered out; 120,719 variants passed the QC criteria, out of which 66,737 had additive transmission in one or more informative families.

To maintain the validity of our analyses, we increased the otherwise low numbers of shared genotyped markers in both studies (175,373 and 9366, respectively) with a multipoint haplotype-based genetic imputation procedure. Phasing was done by SHAPEIT whereas imputation was carried out with IMPUTE by using the PhaseIv3 release of the 1000 genomes as reference. After applying conservative post-imputation QC filters, the numbers of shared markers increased to 6,706,910 and 22,279 in the first and second studies, respectively.

Gene Selection, Statistical Analyses

Three extended candidate gene sets were assembled for the candidate gene study of suicidal behavior. Geneset1 (n=35) was compiled through a literature search, by selecting genes with at least suggestive genome-wide association or expression p-values from studies of suicidal behavior that analyzed independent samples. Geneset2 (n=68) contained microRNA (miRNA) genes expressed in the brain regions essential in executive functions (the dorsolateral and orbitofrontal prefrontal circuits) and genes essential in the miRNA biosynthesis and turnover. Geneset3 extended Geneset2 by including the predicted miRNA target genes that were expressed in the same brain areas (n=11259). Gene expression was evaluated with the Gene Expression Atlas at the European Bioinformatics Institute, whereas miRNA target predictions were extracted from the miRanda and mirSVR databases accessed from microRNA.org.

Association of the genes in Geneset1 with suicidal behavior was assessed with three series of region-based analyses on the MDD, bipolar, and combined datasets. To reduce bias stemming from between-sample heterogeneity, datasets were combined in a binary effect meta-analysis, a modified random-effect meta-analysis implemented in METASOFT. Region-based associations were computed with the hybrid set-based test (HYST) implemented in the KGG analysis suite. The HYST-test computes the region-based statistics from the SNP-based associations based on the underlying linkage values, therefore combining the advantages of the “best SNP” and “SNP combination” approaches. The SNP-based associations were calculated with PLINK, whereas the n=379

European samples of the 1000 Genomes Phase1v3 release were used as a linkage disequilibrium reference. Gene boundaries were set according to refSeq extended by 5 kb at both ends. P-values were corrected according to an α=0.05 family-wise error rate; the two analysis families contained the independent region-based hypotheses (1: MDD + bipolar, and 2: combined analysis). Heritability explained by the analyzed gene-regions were computed with GCTA to provide an effect-size measure for the region-based associations.

In addition, we tested the predictive power of the polygenic risk scores derived from all genesets. Polygenic predictions were assessed with two different methods: PRSice, a standard "pruning and thresholding" approach, and LDpred, a Bayesian posterior PRS prediction test allowing for calibration and maximizing information. Both tests were run using default options by with the MDD sample as a training set and bipolar sample as a target dataset.

Our second study on genetic associations of DSM-IV schizophrenia employed a two-stage analytic design with a hypothesis-generating discovery step and a hypothesis-testing validation step. The discovery analyses were conducted on the SCHIZOBANK trio data and the suggestive associations (p<0.1) were validated with the merged schizophrenia GWAS samples. We used KGG to estimate the region-based gene (extended Simes'-test, GATES), the canonical pathway and positional geneset associations (HYST), and the canonical pathway and positional geneset enrichment analyses (Wilcoxon signed rank-sum test). Gene boundaries were set according to the GENCODE v19 coordinates extended by 5kb in both ends. For downstream multilocus analyses, SNP-based summary association statistics were computed by using a whole-exome Family-Based Association Test (FBAT). An additive inheritance model, at least 4 informative families, and adaptive permutation in the discovery set, and additive logistic regression models controlled for age, sex and population stratification in the validation samples were included. Population stratification was corrected via principal component analysis (PCA) method.

In addition to the multilocus tests of KGG, we performed a functional annotation clustering with DAVID as well. We used the same combined, functional prioritization algorithm as with the suggestive SNP associations to select the input gene lists in both datasets. Clusters with enrichment score (ES) > 1.25 were considered suggestive and were tested for replication during the validation step.

P-values of the validation stage were corrected per familywise error rate α=0.05 for each analysis family (gene-based association, geneset association, geneset enrichment, and functional annotation cluster enrichment). Replication probability was also computed for all multilocus methods based on the corrected p-vales.

Results

Gene-Based Associations with Suicidal Behavior

Gene-based analyses yielded several nominally significant associations, but only DICER1 gene was associated with suicidal behavior with at least nominal significance in the MDD, bipolar, and the combined analyses. In the GAIN bipolar sample, only nominally significant associations were detected between suicidal behavior and genes hsa-miR-195 ($p=0.017$, $h^2=0.003$), CD44 ($p=0.019$, $h^2=0.009$) and DICER1 ($p=0.034$, $h^2=0.004$). Among subjects with MDD, the NTRK2 gene had the strongest association with suicidal behavior ($p=0.0002$, $h^2=0.01$), followed by NXPH1 ($p=0.012$, $h^2=0.014$), GRIA3 ($p=0.017$, $h^2=0.016$), DICER1 ($p=0.032$, $h^2=0.006$) and SPTLC1 ($p=0.041$, $h^2=0.006$). After correction for multiple comparisons, only the association between the NTRK2 gene and suicidal behavior in the MDD sample remained significant ($p_{corr}=0.014$).

Polygenic Risk Prediction of Suicidal Behavior

Among the three genesets, only the PRS scores derived from the miRNA expression geneset (Geneset2) in the MDD sample predicted suicidal behavior with at least nominal significance, albeit with negative regression coefficients in the bipolar dataset. The PRSice analysis with $p_t<0.03$ yielded the largest explained phenotypic variance (Nagelkerke $R^2=0.01$, $p<0.007$). The LDpred analysis indicated the posterior PRS scores with best calibration and accuracy at $p<0.01$ fraction of causal in

Geneset2, predicting suicidal behavior with p=0.019, Nagelkerke R^2=0.0076 and beta=-1.52 in the bipolar sample.

Gene- and Geneset-Based Associations with Schizophrenia

The two-staged analysis of schizophrenia did not indicate any genes or canonical pathways with significant p-values after correction for multiple comparisons. However, among the positional genesets, the 14q31, 5q31 and Xq13 cytobands had significant association or enrichment p-values after correction for multiplicity ($p_{\_corr}$: 0.002, 0.006, and 0.048 respectively).

Functional Annotation Clustering in Schizophrenia

Functional annotation clustering resulted in several nominally significant clusters in both samples (ES > 1.3). Among those, clusters of splicing/alternative splicing (ES_{exp}: 2.85, ES_{rep}: 23.93), brain development (ES_{exp}: 1.23, ES_{rep}: 4.55) and embryonic development (ES_{exp}: 1.34, ES_{rep}: 4.77) had ES scores corresponding to a p-value corrected for multiple comparisons. Replication probability increased in parallel with the complexity of the multilocus method ($P_{\_rep}$ for gene, geneset, functional annotation cluster: 0, 0.015, and 0.21)

Conclusions

- Cross-disorder, region-based analysis of suicidal behavior highlighted the importance of the BDNF-NTRK2-CREB pathway in suicidal ideation
- The nominally significant, cross-disorder association of the DICER1 gene and the PRS predictions suggest a potential mediating role of the miRNA system. However, the lack of other cross-disorder associations and the inconsistent predictions indicate that suicidal ideation and attempts are two distinct phenotypes.
- The high proportion of nominally significant associations annotated as ENCODE regulatory loci, affirm their validity and highlight the necessity of integrating genomic, epigenomic and other functional annotations to make proper interpretations.
- Results of our second study implicated the involvement of neurodevelopment, synaptic plasticity, and the immune system in the etiology of schizophrenia, further highlighting the importance of integrating functional annotations to data analysis.
- Our findings suggest improved replication probability of multilocus methods, even in cases with large genetic heterogeneity, but not for those with definite phenotypic heterogeneity.

List of Publications

Publications related to this thesis:

1. **Pulay AJ**, Rethelyi JM. (2016) Multimarker analysis suggests the involvement of BDNF signaling and microRNA biosynthesis in suicidal behavior. *AMERICAN JOURNAL OF MEDICAL GENETICS PART B-NEUROPSYCHIATRIC GENETICS* 171:(6) pp. 763-776.

2. **Pulay AJ**, Koller J, Nagy L, Molnár MJ, Réthelyi J, Magyar SCHIZOBANK Konzorcium munkatársai. (2017) A szkizofrénia multilókusz genetikai vizsgálata az idegfejlődés és az immunrendszer zavarának oki szerepére utal(hat). *IDEGGYOGYASZATI SZEMLE-CLINICAL NEUROSCIENCE* 70:(3-4) pp. 115-126.

Other publications:

1. Benkovits J, Magyarosi S, **Pulay AJ**, Makkos Z, Egerhazi A, Balogh N, Almos P, Liko I, Schizobank Consortium H, Nemeth G, Molnar JM, Nagy L, Rethelyi JM- (2016) CNTF, COMT, DDR1, DISC1, DRD2, DRD3 es DTNBP1 kandidáns gének vizsgálata szkizofréniában: eredmények a Magyar SCHIZOBANK Konzorcium vizsgálatából. *NEUROPSYCHOPHARMACOLOGIA HUNGARICA* 18**:**(4) pp. 181-187.

2. **Pulay AJ**, Bitter I, Papp S, Gulácsi L, Péntek M, Brodszky V, Hevér NV, Rencz F, Baji P. (2016) Exploring the Relationship between Quality of Life (EQ-5D) and Clinical Measures in Adult Attention

Deficit Hyperactivity Disorder (ADHD) *APPLIED RESEARCH IN QUALITY OF LIFE* First online: 20 April 2016: pp. 1-16.

3. Szkultecka-Debek M, Walczak J, Augustynska J, Miernik K, Stelmachowski J, Pieniazek I, Obrzut G, Pogroszewska A, Paulic G, Damir M, Antolic S, Tavcar R, Indrikson A, Aadamsoo K, Jankovic S, **Pulay AJ**, Rimay J, Varga M, Sulkova I, Verzun P. (2015) Epidemiology and Treatment Guidelines of Negative Symptoms in Schizo-phrenia in Central and Eastern Europe: A Literature Review. *CLINICAL PRACTICE AND EPIDEMIOLOGY IN MENTAL HEALTH* 11: pp. 158-165.

4. Kerridge BT, Saha TD, Smith S, Chou PS, Pickering RP, Huang B, Ruan JW, **Pulay AJ.** (2011) Dimensionality of hallucinogen and inhalant/solvent abuse and dependence criteria: implications for the diagnostic and statistical manual of mental disorders-Fifth edition. *ADDICTIVE BEHAVIORS* 36:(9) pp. 912-918.

5. Pickering RP, Goldstein RB, Hasin DS, Blanco C, Smith SM, Huang B, **Pulay AJ**, Ruan WJ, Saha TD, Stinson FS, Dawson DA, Chou SP, Grant BF. (2011) Temporal relationships between overweight and obesity and DSM-IV substance use, mood, and anxiety disorders: results from a prospective study, the National Epidemiologic Survey on Alcohol and Related Conditions. *JOURNAL OF CLINICAL PSYCHIATRY* 72:(11) pp. 1492-1502.

6. Dawson DA, **Pulay AJ**, Grant BF. (2010) A comparison of two single-item screeners for hazardous drinking and alcohol use disorder. *ALCOHOLISM-CLINICAL AND EXPERIMENTAL RESEARCH* 34:(2) pp. 364-374.

7. **Pulay AJ**, Stinson FS, Ruan WJ, Smith SM, Pickering RP, Dawson DA, Grant BF. (2010) The relationship of DSM-IV personality disorders to nicotine dependence-results from a national survey. *DRUG AND ALCOHOL DEPENDENCE* 108:(1-2) pp. 141-145.

8. Saha TD, Compton WM, **Pulay AJ**, Stinson FS, Ruan WJ, Smith SM, Grant BF. (2010) Dimensionality of DSM-IV nicotine dependence in a national sample: an item response theory application. *DRUG AND ALCOHOL DEPENDENCE* 108:(1-2) pp. 21-28.

9. Srivastava V, Buzas B, Momenan R, Oroszi G, **Pulay AJ**, Enoch MA, Hommer DW, Goldman D. (2010) Association of SOD2, a mitochondrial antioxidant enzyme, with gray matter volume shrinkage in alcoholics. *NEUROPSYCHOPHARMACOLOGY* 35:(5) pp. 1120-1128.

10. Grant BF, Goldstein RB, Chou SP, Huang B, Stinson FS, Dawson DA, Saha TD, Smith SM, **Pulay AJ**, Pickering RP, Ruan WJ, Compton WM. (2009) Sociodemographic and psychopathologic predictors of first incidence of DSM-IV substance use, mood and anxiety disorders: results from the Wave 2 National Epidemiologic Survey on Alcohol

and Related Conditions. *MOLECULAR PSYCHIATRY* 14:(11) pp. 1051-1066.

11. **Pulay AJ**, Stinson FS, Dawson DA, Goldstein RB, Chou SP, Huang B, Saha TD, Smith SM, Pickering RP, Ruan WJ, Hasin DS, Grant BF. (2009) Prevalence, Correlates, Disability, and Comorbidity of DSM-IV Schizotypal Personality Disorder: Results From the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *PRIMARY CARE COMPANION TO THE JOURNAL OF CLINICAL PSYCHIATRY* 11:(2) pp. 53-67.

12. Grant BF, Chou SP, Goldstein RB, Huang B, Stinson FS, Saha TD, Smith SM, Dawson DA, **Pulay AJ**, Pickering RP, Ruan WJ. (2008) Prevalence, correlates, disability, and comorbidity of DSM-IV borderline personality disorder: results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *JOURNAL OF CLINICAL PSYCHIATRY* 69:(4) pp. 533-545.

13. Gundy S, Szekely G, Farkas G, **Pulay A**, Remenar E. (2008) Biomarkerek alkalmazása során felmerülő problémák malignus és nem malignus betegségben szenvedő alkoholisták esetében: Problems occurring in the application of cytogenetic biomarkers for alcoholics with and without malignant diseases. *MAGYAR ONKOLÓGIA* 52:(2) pp. 153-161.

14. **Pulay AJ**, Dawson DA, Ruan WJ, Pickering RP, Huang B, Chou SP, Grant BF. (2008) The relationship of impairment to personality disorder severity among individuals with specific axis I disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *JOURNAL OF PERSONALITY DISORDERS* 22:(4) pp. 405-417.

15. **Pulay AJ**, Dawson DA, Hasin DS, Goldstein RB, Ruan WJ, Pickering RP, Huang B, Chou SP, Grant BF. (2008) Violent behavior and DSM-IV psychiatric disorders: results from the national epidemiologic survey on alcohol and related conditions. *JOURNAL OF CLINICAL PSYCHIATRY* 69:(1) pp. 12-22.

16. Stinson FS, Dawson DA, Goldstein RB, Chou SP, Huang B, Smith SM, Ruan WJ, **Pulay AJ**, Saha TD, Pickering RP, Grant BF. (2008) Prevalence, correlates, disability, and comorbidity of DSM-IV narcissistic personality disorder: results from the wave 2 national epidemiologic survey on alcohol and related conditions. *JOURNAL OF CLINICAL PSYCHIATRY* 69:(7) pp. 1033-1045.

17. Goldstein RB, Compton WM, **Pulay AJ**, Ruan WJ, Pickering RP, Stinson FS, Grant BF. (2007) Antisocial behavioral syndromes and DSM-IV drug use disorders in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *DRUG AND ALCOHOL DEPENDENCE* 90:(2-3) pp. 145-158.