

Current Practices in Genetics Research of Psychiatric Disorders

Chunyu Liu^{1,2*}, Kay S. Grennan², Elliot S. Gershon³

1. State Key Laboratory of Medical Genetics, Central South University, Changsha, China;

2. Department of Psychiatry, University of Illinois at Chicago, Chicago, USA;

3. Department of Psychiatry and Behavioral Neuroscience, The University of Chicago, Chicago, USA

***Correspondence:** Dr. Chunyu Liu, email: liucy@uic.edu, Department of Psychiatry University of Illinois at Chicago

ABSTRACT

After a brief review of major factors that influence results of genetic studies, we summarize key considerations in designing genetic studies of psychiatric disorders. A sufficient sample size is needed, along with an appropriate threshold for statistical significance. Rigorous quality assessment, controls of covariates and population structure, and replication of results in a second data set are critical to avoid spurious findings. Data sharing should be promoted to enable powerful big data analyses. Gradually, functional studies will be incorporated into genetic studies to improve the quality of genetic studies and biological interpretation of statistical associations.

INTRODUCTION

Incorporating a genetic component can be appealing to an investigator designing a new study of a neuropsychiatric disorder. After all, it is well known that most neuropsychiatric disorders are heritable to one degree or another. Incorporating a genetic component into a study may seem straightforward: draw blood from subjects, submit it to a service lab, get the genotypes back, and test them for an association with the phenotype(s) being studied. In reality, it is considerably more complicated than that. Putting aside the complexity of choosing the proper platforms for genotyping and phenotyping, there are a number of issues that must be addressed in the initial study design if the results are to be replicable and relevant to the current state of knowledge of the disorder in question. The three major issues are effect sizes, Type I error rates, and genetic and phenotypic heterogeneity.

The effect size of a particular gene variant can come as a bit of surprise to researchers unfamiliar with genetics of complex traits or common disorders. It is a major feature that sets common disorders apart from rare, Mendelian disorders. For a disease with prevalence of roughly 1%, a measure of effect size is the ratio of variant frequency in patients to that in controls. The single gene variants for rare diseases such as Huntington's disease typically have effect sizes greater than 1000, but almost all discovered common variant associations for psychiatric disorders have effect sizes between 1.1 and 1.4 [1]. Since sample size and effect size are the primary determinants of statistical power, genome-wide association study (GWAS) of complex disorders,



OPEN ACCESS

DOI: 10.20900/jpbs.20160001

Received: January 10, 2016

Accepted: February 6, 2016

Published: April 25, 2016

website: <http://jpbs.qingres.com>

Copyright: ©2016 Cain et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

including neuropsychiatric disorders, require many thousands of samples in a case-control study to produce any genome-wide significant associations.

Here, “genome-wide significant associations” indicates that the association p-value has been corrected for multiple testing, a correction which controls for Type I error. When performing as many tests as are required for a genome-wide association test, the multiple testing burden becomes a major problem. For example, testing 10,000 loci for association with a phenotype, using the standard p-value cutoff of 0.05, will produce 500 false positives, or Type I errors, without some type of statistical correction. Type I error has been the primary concern in small sample studies with low allele frequency variants, which tend to give spurious associations.

Heterogeneity refers to diversity among samples. In neuropsychiatric studies, diversity among the disorder-associated traits present in the case subjects is a particular concern[2]. Clinicians well understand that a single psychiatric diagnosis can cover a broad range of cognitive function or mood changes; problems with psychiatric nosology have been discussed for over a hundred years[3]. This phenotypic heterogeneity among cases is likely one of the reasons that the results of psychiatric case-control studies are often inconclusive or hard to replicate. Efforts are being made to bring clinical diagnostic categories into line with the underlying biology, via new phenotyping constructs such as those in the NIMH’s Research Domain Criteria (RDoC)[4]. Any screening that can reduce sample heterogeneity would lead to better power in detecting genetic factors[5]. There has been little success so far, however, so the problem of phenotypic heterogeneity, either within diagnostic categories or across traits like behavior or imaging, must be kept in mind when designing neuropsychiatric association studies.

Case samples can also be, and probably are, genetically heterogeneous, meaning that what appears to be a single phenotype can actually be associated with any one of multiple genetic variants. For example, in psychiatry, we know that almost 30% of adults with the rare 22q11.2 deletion are diagnosed with schizophrenia[6]. However, the vast majority (99%) of schizophrenia patients do not have this deletion[7], so their disease must be attributable to other genetic and environmental factors (although there could also be common effects of genes within the fairly large 22q11.2 deletion). Schizophrenia is therefore a genetically heterogeneous disorder. Different mutations of different genes can lead to similar clinical phenotypes and diagnoses. Meanwhile, the same mutation could present different clinical features in different individuals likely due to genetic background differences, which brings us back to the problem of phenotypic heterogeneity.

Genetic heterogeneity can also refer to population stratification, which is a problem in GWASs in general, not just for neuropsychiatric conditions[8]. Population stratification can produce spurious association results when there are allele frequencies between ethnic groups that are unrelated to disease, so a statistical correction must be incorporated into association study designs.

What follows is a checklist for taking the above issues into account when designing or performing a psychiatric genetic study.

1. Make Sure You Have the Right Samples for the Kind of Study You Want to Perform. There are many choices in designing genetic studies, from family-based to case-control, from genome-wide to candidate gene, and from linkage to association. Selection of methods depends on the specific hypothesis being tested. Detection of rare de novo mutations requires nuclear families. Association tests of common variants work better with large case-control cohorts. Private large effect mutations may be detected through large familial samples. Each type of study requires a particular statistical method, each of which comes with its own null hypothesis and basic assumptions.

2. Perform a Power Analysis. The number of subjects recruited into the study should be based on the expected effect sizes for the corresponding genetic variants and phenotypes in the study (as explained above, they will probably be quite small). One should be prepared to have 80% power to detect the effect at a given Type I error rate, taking the multiple testing correction into account. It should be noted that rare copy number variants and functional de novo mutations are predicted to have larger effect sizes[9]; therefore, studies with smaller sample sizes can be useful, as well. Strong effect rare events can be detected with smaller sample sizes than are needed for GWAS using algorithms to collapse the different events into summary statistics, like gene-wide or region-wide tests[10].

3. Select the Appropriate Threshold of Significance. Because of the multiple testing burden, GWASs typically use a p-value of $2e-8$ or $5e-8$ as a threshold for genome-wide significance. If a study focuses on some candidate variants or genes, the number of linkage equilibrium (LD) independent tests could be used to adjust that threshold. False discovery rate (FDR)[11], Bonferroni correction, or permutation are a few

commonly used methods for correcting multiple testing inflation of Type I error rates.

4. Perform Rigorous Quality Control (QC) Measures. Insufficient quality control can lead to both false positives and false negatives. Both sample and measurement quality should be carefully assessed. Different experimental platforms have their own specific quality evaluation procedures; for example, a microarray-based genotyping should be evaluated for call rates for each sample and each SNP. Proper filtering should be applied to remove low-quality data before association tests are performed. Read-depth and sequence quality should be assessed for sequencing-based genotype data. The Hardy-Weinberg equilibrium test is frequently used to identify some potential genotyping errors. Because of the potential for sample handling errors, identity checks and sex checks are also frequently used to capture errors.

5. Control for the Effects of Co-variates. When the effects of co-variates are not controlled for statistically, they can skew the association results to produce false positives or false negatives. Many biological or environmental factors can be confounded with case-control status, such as sex, age, and drug exposure. There are also many technical artifacts that can interfere with results, such as experimental batch effects; these are the differences that arise between experimental batches due to anything from different reagent lots to different technicians running the experiments to variation in atmospheric pressure on the days the experiments were run[12]. There are programs designed to statistically remove these effects from data, such as COMBAT[13], which removes batch effects from gene expression microarray data. These programs are not a cure-all, so a careful plan for randomized placement of samples in experiments is still required.

6. Account for Population Stratification in Case-control Studies. Ensuring that the case and control groups have comparable population composition is necessary to avoid false positives. Software like STRUCTURE[14] or methods like principal component analysis (PCA) are frequently used to address the problem using unlinked or ancestry informative markers (AIMs) that represent different geographical origins. These methods for controlling population stratification are better used when the stratification is mild. When samples come from very different populations, like from different continental groups, it may be better to study the different population samples separately, followed by a combined statistical analysis.

7. Include Validations, i.e., Replication of Results, in Study Design and Budget. Validation of any positive results is essential. It is not enough to observe a significant result in only one data set. Ideally, an appropriate second data set will be available with which to re-run the analysis and confirm the findings. The significance level of the initial findings can be an indicator of the probability that the results will be reproduced. However, the results may not be reproduced in a second data set, due to either false positives in the initial study or sample heterogeneity problems. Note that the significance threshold for replication can be much lower than in the initial discovery, which is typically genome-wide, because only the loci with significant associations in the initial data will be tested in the replicate data.

8. Share Data. Since sample size and data quality is critical for the success of genetic studies, data sharing is becoming the trend in the research community, due in part to support from funding agencies and publishers. The NIH now mandates data sharing for large grants with genetics and genomics data. The Psychiatric Genomics Consortium (PGC, <http://www.med.unc.edu/pgc/>) and Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA, <http://enigma.ini.usc.edu/>) have set good examples.

9. Remember that Proof of Biological Function is Needed to Substantiate Genetic Statistical Associations. Much, though not all, of the low-hanging fruit in genetic association studies have already been plucked. The field has moved past looking for simple Mendelian factors, and towards identifying functional aspects of the molecular changes underlying challenging complex disorders. Today, pathway analysis, network analysis[15, 16] and in silico prediction of functional variation produced by coding variants are common practice. Functional analysis of non-coding variants is a major challenge facing the field[17]. Expression quantitative trait loci (eQTL) mapping is one way to address the issue, since it identifies variants with potential roles in regulating gene expression[18]. Association with other phenotypes like brain imaging traits is another way of investigating possible functional effects of genetic variants[19, 20]. As more genomic and epigenomic data accumulate and more effort is put into studies of non-coding variants, particularly for human brains[21], we should acquire more and more evidence as to the functional effects of the genetic variants identified by genetic association studies. Converting statistical associations into biological causalities will be the major task after genetic studies.

To summarize, investigators must be aware that heterogeneity is a particular problem in genetic studies of neuropsychiatric diseases: what is referred to as a single diagnosis can actually represent a range of phenotypic traits, while on the flip side a single phenotypic trait can be produced by multiple alleles. Study design must include sufficient power to detect very small effect sizes, proper quality and covariate control procedures, a validation step, and a significance threshold that will control for Type I error. Ideally, studies

will produce functional evidence to support identified statistical associations between genetic variants and phenotypes. Data sharing should be promoted as a common practice of genetic studies.

REFERENCES

1. Vinkhuyzen AA, Wray NR. Novel directions for G x E analysis in psychiatry. *Epidemiol Psychiatr Sci*. 2015;24(1):12-9. doi: 10.1017/S2045796014000584. PubMed PMID: 25295644; PubMed Central PMCID:PMC4411743.
2. Chaste P, Klei L, Sanders SJ, Hus V, Murtha MT, Lowe JK, et al. A genome-wide association study of autism using the Simons Simplex Collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biological psychiatry*.2015;77(9):775-84. doi: 10.1016/j.biopsych.2014.09.017. PubMed PMID: 25534755; PubMed Central PMCID: PMC4379124.
3. Gershon ES, Grennan KS. Genetic and genomic analyses as a basis for new diagnostic nosologies. *Dialogues in clinical neuroscience*. 2015;17(1):69-78. Epub 2015/05/20. PubMed PMID: 25987865; PubMed Central PMCID: PMC4421903.
4. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*. 2013;11:126. Epub 2013/05/16. doi: 10.1186/1741-7015-11-126. PubMed PMID: 23672542; PubMed Central PMCID: PMC3653747.
5. Hansen WB, Collins LM. Seven ways to increase power without increasing N. *NIDA Res Monogr*. 1994;142:184-95. PubMed PMID: 9243537.
6. Monks S, Niarchou M, Davies AR, Walters JT, Williams N, Owen MJ, et al. Further evidence for high rates of schizophrenia in 22q11.2 deletion syndrome. *Schizophrenia research*. 2014;153(1-3):231-6. Epub 2014/02/19. doi: 10.1016/j.schres.2014.01.020. PubMed PMID: 24534796.
7. Bassett AS, Chow EW. Schizophrenia and 22q11.2 deletion syndrome. *Current psychiatry reports*. 2008;10(2):148-57. Epub 2008/05/14. PubMed PMID: 18474208; PubMed Central PMCID: PMC3129332.
8. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nature reviews Genetics*. 2010;11(7):459-63. Epub 2010/06/16. doi: 10.1038/nrg2813. PubMed PMID: 20548291; PubMed Central PMCID: PMC2975875.
9. Thapar A, Cooper M. Copy number variation: what is it and what has it told us about child psychiatric disorders? *J Am Acad Child Adolesc Psychiatry*. 2013;52(8):772-4. doi: 10.1016/j.jaac.2013.05.013. PubMed PMID:23880486;PubMed Central PMCID: PMC3919207.
10. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5-23. doi: 10.1016/j.ajhg.2014.06.009. PubMed PMID: 24995866; PubMed Central PMCID: PMC4085641.
11. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125(1-2):279-84. PubMed PMID: 11682119.
12. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*. 2011;6(2):e17238. doi: 10.1371/journal.pone.0017238. PubMed PMID: 21386892; PubMed Central PMCID: PMC3046121.
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-27.
14. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2):945-59. PubMed PMID: 10835412; PubMed Central PMCID: PMC1461096.
15. Grennan KS, Chen C, Gershon ES, Liu C. Molecular network analysis enhances understanding of the biology of mental disorders. *Bioessays*. 2014;36(6):606-16. Epub 2014 Apr 14. doi: 10.1002/bies.201300147. PubMed PMID: 24733456;PubMed Central PMCID: PMC4300946.
16. Chen C, Cheng L, Grennan K, Pibiri F, Zhang C, Badner JA, et al. Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry*. 2013;18(12):1308-14. Epub 2012/11/14. doi: 10.1038/mp.2012.146. PubMed PMID: 23147385; PubMed Central PMCID: PMC4018461.

17. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics*. 2013;93(5):779-97. doi: 10.1016/j.ajhg.2013.10.012. PubMed PMID: 24210251; PubMed Central PMCID: PMC3824120.
18. Liu C. Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases. *Neurosci Bull*. 2011;27(2):123-33.
19. Li M, Wang Y, Zheng XB, Ikeda M, Iwata N, Luo XJ, et al. Meta-analysis and brain imaging data support the involvement of VRK2 (rs2312147) in schizophrenia susceptibility. *Schizophr Res*. 2012;142(1-3):200-5. doi: 10.1016/j.schres.2012.10.008. PubMed PMID: 23102693.
20. Potkin SG, Macciardi F, Guffanti G, Fallon JH, Wang Q, Turner JA, et al. Identifying gene regulatory networks in schizophrenia. *Neuroimage*. 2010;53(3):839-47. doi: 10.1016/j.neuroimage.2010.06.036. PubMed PMID: 20600988; PubMed Central PMCID: PMC3055795.
21. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci*. 2014;17(6):782-90. doi: 10.1038/nn.3708. PubMed PMID: 24866044; PubMed Central PMCID: PMC34112149.