EDITORIAL

AMERICAN JOURNAL OF medical genetics PART B
Neuropsychiatric Genetics

The *P*-Value Illusion: How to Improve (Psychiatric) Genetic Studies

Alexander B. Niculescu[1,2]* and Helen Le-Niculescu[1]

[1]Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana

[2]Indianapolis VA Medical Center, Indianapolis, Indiana

Received 29 December 2009; Accepted 5 February 2010

There is an emerging appreciation that genome-wide association studies (GWAS) have failed to live up to expectations and deliver major advances to date. A "surge" strategy, of pooling resources and increasing number of subjects tested, is underway. We argue that, while useful, it will not be enough by itself. Complementary approaches are needed to mine these large datasets. We describe a series of problems, opportunities, and offer a potential comprehensive solution. © 2010 Wiley-Liss, Inc.

Key words: **genome-wide association studies (GWAS); convergent functional genomics (CFG); copy-number variants (CNV); psychiatric genetics; complexity**

How to Cite this Article:
Niculescu AB, Le-Niculescu H. 2010. The *P*-Value Illusion: How to Improve (Psychiatric) Genetic Studies.
Am J Med Genet Part B 153B:847–849.

COMPLEXITY

The complexity of most medical and psychiatric disorders as currently defined is such that a large repertoire of genes is likely involved [Le-Niculescu, 2009a; Manolio, 2009]. One way to make things more tractable is to use more narrow phenotypes (sub-phenotypes) for genetic and biomarker studies. Additionally, using a case–case design, comparing different sub-phenotypes or extremes in quantitative sub-phenotypes, may reduce noise and increase power [Kurian, 2009; Le-Niculescu, 2009b], compared to case–control approaches.

HETEROGENEITY

By and large, a small number of housekeeping-type genes have been identified to date by genome-wide association studies (GWAS) [Anon, 2007; Baum, 2008], as opposed to the more numerous, interesting and biologically relevant genes implicated by other approaches. It may be that the GWAS are under-powered, and can only detect genes in which there is a high degree of conservation (low degree of heterogeneity) in the population, such as housekeeping genes [She, 2009]. Genes that are involved in more discrete and evolved functions, such as specific tissue-enriched genes (e.g., brain-enriched genes like BDNF [Liu, 2009; Petryshen, 2009]), may have a higher degree of heterogeneity due to the evolutionary pressures of adapting to the environment. A mutation in a highly conserved gene will stand out, have a higher impact, and be detectable by GWAS. A mutation in a highly heterogeneous gene will not stand out across the diverse repertoire of common variants, and will be harder to detect with GWAS. Such interesting and relevant genes, however, may be identified by gene expression studies and other biological experiments, in human tissue samples and animal models.

Copy Number Variants (CNVs) may arise all the time, likely as a paternal (or maternal) age effect of cumulative damage to germline DNA. These random transgenic effects (over-expression, knock-outs) are inadvertent experiments that nature did for us. They are an additional and previously unsuspected source of diversity and heterogeneity in the population, resulting in illness as well as possibly in supra-normative individuals. A note of caution is that studies looking at CNVs need to be carried out on DNA obtained from non-immortalized cell lines, to prevent cell culture and viral (e.g., Ebstein–Barr virus in lymphoblastoid cell lines) induced artifacts. More generally, the tissue source of DNA is relevant, and generalizing from a single one is dangerous, as there is an emerging appreciation of mozaicism and DNA heterogeneity among neural and non-neural cell types [Coufal, 2009].

Conflict of interest: ABN is a scientific co-founder of Mindscape Diagnostics, Inc.

*Correspondence to:
Alexander B. Niculescu, M.D., Ph.D., Assistant Professor of Psychiatry and Medical Neuroscience, Indiana University School of Medicine, Staff Psychiatrist, Indianapolis VA Medical Center Director, INBRAIN and Laboratory of Neurophenomics, Institute of Psychiatric Research, 791 Union Drive, Indianapolis, IN 46202-4887. E-mail: anicules@iupui.edu

Published online 17 March 2010 in Wiley InterScience (www.interscience.wiley.com)
DOI 10.1002/ajmg.b.31076

Next generation sequencing, where the whole genome will be sequenced in increasing number of individuals, will reveal rather than solve the problem of heterogeneity. In lieu of just awaiting the hypothetical development of new bioinformatic approaches to deal with that expected torrent of data, sensible strategies feasible with current knowledge and means should be pursued, on the existing and emerging data.

REPRODUCIBILITY

We would like to suggest that *P*-values at best permit the precise ranking of findings *within* a study. At worst, they are over-interpreted and a source of frustration in terms of reproducibility of findings *across* studies. This ''*P*-value illusion'' in genetics, that a significant *P*-value in one study should necessarily reproduce in another independent study, is based on an under-appreciation of two factors. The first one is the fit-to-cohort effect of classic statistical analyses of genetic studies, the second one is the above discussed complexity and heterogeneity of most disorders. In essence, you are identifying the genes that are a best fit in the cohort you derive your data from. This guarantees that your top findings will not be at the top of the list in a study carried in a different cohort, since complexity and heterogeneity will ensure that no two cohorts are alike. In addition to population-specific factors, which we think are the primary reason, there may be experimental stochastic factors involved as well. This phenomenon is sometimes describes as the ''winner's curse,'' a strong initial finding not being as strong in subsequent independent cohorts.

A solution is to use a fit-to-disease approach, like Convergent Functional Genomics (CFG) [Niculescu, 2000; Ogden, 2004; Le-Niculescu, 2007, 2009a,b; Rodd, 2007; Kurian, 2009]. Such an approach, in addition to *P*-values, uses multiple independent lines of evidence related to illness, including gene expression studies, as a way of prioritizing findings within a cohort, similar to a Google PageRank algorithm. Genes prioritized in such a way may not have the highest *P*-values, but will generalize and reproduce well in independent cohorts. Since this approach is based on a gene level integration of data rather than a SNP level analysis, it reduces heterogeneity. It can and has been used profitably to mine GWAS [Le-Niculescu, 2009a] and biomarker [Kurian, 2009; Le-Niculescu, 2009b] datasets, and to extract panels of top genes or biomarkers that reproduce well in independent cohorts. Studies that by themselves are relatively under-powered can be mined and made to yield results using CFG, by bringing to bear other large datasets and databases relevant to that disease, resulting in essence in a field-wide collaboration. Cross-validating signals from other sources, with different noise factors, can increase the signal–noise ratio and decrease the size of cohort (n) you need.

HERITABILITY

Heritability estimates based on small twin studies may be overly optimistic [Kaminsky, 2009]. Additionally, the cumulative combinatorics of common gene variants and environment (CCxCGV XE) [Niculescu, 2010; Patel, 2010] may play a more prominent role than anticipated in complex disorders [Manolio, 2009], such as psychiatric disorders [Le-Niculescu, 2009a].

CONCLUSIONS

We suggest that the large and continuously growing databases from GWAS are a very useful resource for breaking the genetic code of complex disorders, such as psychiatric disorders. Combining them with complementary approaches that (1) look at sub-phenotypes, (2) use a case–case design, (3) integrate and prioritize findings at a gene level using multiple lines of evidence, including gene expression, has a better chance of yielding findings that are disease relevant, as well as reproducible and predictive in independent cohorts. That is the key litmus test, in our opinion, for any biomarker (or genetic) study. For a complete understanding of the illness, the analyses then need to be pursued at a biological pathway and mechanistic level, integrating environmental effects as key modulators of gene expression and phenotype manifestation. A (r)evolution in medical nosology in general, and psychiatric nosology in particular [Kendler et al., 2010; Niculescu et al., 2009], may follow.

ACKNOWLEDGMENTS

We would like to thank Martin Schalling and Dan Koller for useful discussions. This work was supported by a VA Merit Award to ABN.

REFERENCES

Anon. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Baum AE, et al. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry 13:197–207.

Coufal NG, et al. 2009. L1 retrotransposition in human neural progenitor cells. Nature 460:1127–1131.

Kaminsky ZA, et al. 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41:240–245.

Kendler KS, Muñoz RA, Murphy G. 2010. The development of the Feighner criteria: A historical perspective. Am J Psychiatry 167:134–142.

Kurian SM, Le-Niculescu H, Patel SD, Bertram D, Davis J, Dike C, Yehyawi N, Lysaker P, Dustin J, Caligiuri M, Lohr J, Lahiri DK, Nurnberger JI Jr, Faraone SV, Geyer MA, Tsuang MT, Schork NJ, Salomon DR, Niculescu AB. 2009. Identification of blood biomarkers for psychosis using convergent functional genomics. Mol Psychiatry [Epub ahead of print].

Le-Niculescu H, et al. 2007. Towards understanding the schizophrenia code: An expanded convergent functional genomics approach. Am J Med Genet Part B 144B:129–158.

Le-Niculescu H, et al. 2009a. Convergent functional genomics of genome-wide association data for bipolar disorder: Comprehensive identification of candidate genes, pathways and mechanisms. Am J Med Genet Part B 150B:155–181.

Le-Niculescu H, et al. 2009b. Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry 14: 156–174.

Liu X, et al. 2009. Family-based association study between brain-derived neurotrophic factor gene and major depressive disorder of Chinese descent. Psychiatry Res 169:169–172.

Manolio TA, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Niculescu AB III, et al. 2000. Identifying a series of candidate genes for mania and psychosis: A convergent functional genomics approach. Physiol Genomics 4:83–91.

Niculescu AB III, Schork NJ, Salomon DR. 2009. Mindscape: A convergent perspective on life, mind, consciousness and happiness. J Affect Disord [Epub ahead of print].

Ogden CA, et al. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: An expanded convergent functional genomics approach. 2004. Mol Psychiatry 9:1007–1029.

Petryshen TL, Sabeti PC, Aldinger KA, Fry B, Fan JB, Schaffner SF, Waggoner SG, Tahl AR, Sklar P. 2009. Population genetic study of the brain-derived neurotrophic factor (BDNF) gene. Mol Psychiatry [Epub ahead of print].

Rodd ZA, et al. 2007. Candidate genes, pathways and mechanisms for alcoholism: An expanded convergent functional genomics approach. Pharmacogenomics J 7:222–256.

She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen R. 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics 10:269.