Review

Neuro
epidemiology

Neuroepidemiology 2009;33:276–279
DOI: 10.1159/000235639

Received: February 18, 2009
Accepted: June 5, 2009
Published online: August 20, 2009

# The Era of Genomic Epidemiology

Bryan J. Traynor

Neuromuscular Diseases Research Group, Laboratory of Neurogenetics, NIA, and Neurogenetics Branch, NINDS, Bethesda, Md., USA

**Key Words**
Genetics · Genomics · Epidemiology

**Abstract**
The recent revolution in genomics is already having a profound impact on the practice of epidemiology. The purpose of this commentary is to demonstrate how genomics and epidemiology will continue to rely heavily on each other, now and in the future, by illustrating a number of interaction points between these 2 disciplines: (1) the use of genomics to estimate disease heritability; (2) the impact of genomics on analytical study design; (3) how genome-wide data can be employed to effectively overcome residual population stratification arising from selection bias; (4) the importance of genomics as a tool in epidemiological investigation; (5) the importance of epidemiology in the collection of adequately phenotyped samples for genomics studies, and (6) for unraveling the clinical and therapeutic relevance of genetic variants once they are discovered.



## Introduction

Technological advancements that allow the genotyping of several thousands of single-nucleotide polymorphisms across the entire genome in an efficient, high-throughput manner have essentially allowed the genetics research community to foray beyond the realm of rare mendelian conditions into the arena of common diseases. Since the publication of the first genome-wide association study three years ago [1], there has been a flood of such studies detailing the genetic basis of diseases ranging from inflammatory bowel disease to hypertension. The National Human Genome Research Institute catalog lists 184 genome-wide association studies with over a 1,000 SNPs linked to 130 diseases and traits, and this list expands on an almost daily basis [2]. Undoubtedly, advances in genomics and the knowledge that these studies bring are already having a profound influence on the practice of epidemiology, and the interaction between the two fields will continue to be mutually beneficial, especially as epidemiologists develop novel methods to utilize genomic data [3]. This article attempts to describe the symbiotic relationship that has developed between genomics and epidemiology and to illustrate how the two disciplines will continue to rely heavily on each other for their future success. As in all good relationships, there will be give-and-take, but, overall, both fields will benefit.

### *Heritability Estimation*

The reality is that genomics will overshadow the role of family studies in estimating disease heritability. By its very nature, traditional family study methodology focus-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com



Accessible online at:
www.karger.com/ned

Bryan J. Traynor
Laboratory of Neurogenetics, NIA, NIH, Building 35, Room 1A/1000
35 Convent Drive
Bethesda, MD 20892-3720 (USA)
Tel. +1 301 451 7606, Fax +1 301 451 7295, E-Mail traynorb@mail.nih.gov

es on the collection of family history for a large number of cases and controls, a process often requiring many years to complete. Furthermore, even the most assiduous collection cannot avoid bias arising from incomplete data due to family members not knowing, or being unwilling to provide, pedigree data. Stricter privacy laws are eroding investigators' ability to collect clinical and demographic data on relatives without their knowledge or permission, making family studies even harder to successfully complete.

In contrast, the high density of genotype data generated in genome-wide association studies allows disease heritability to be more accurately and more easily estimated. This approach is not hampered by lack of family history, and the analysis can be completed in a relatively short period of time. Despite the upfront high cost of genome-wide studies, overall they still represent considerable savings compared to longer-term family studies. Even the usefulness of family studies for identifying novel relationships between diseases will eventually be superseded as the genetic architecture of human illness is more fully understood, and pathway analysis (i.e. determining which pathways are perturbed in disease based on genome-wide data) is applied to link apparently disparate diseases. Indeed, genomic-based pathway analysis may serve as the basis of a new classification system of human pathology by grouping diseases arising from defects in the same biological pathways. For example, genome-wide association studies have found that variants in two genes associated with increased risk of diabetes also influence prostate cancer susceptibility among men [4–6]. Pathway analysis of neurodegenerative diseases, such as amyotrophic lateral sclerosis and Parkinson's disease is already attempting to tease apart the cellular mechanisms involved in neuronal cell death [7]. Though such system biology studies should be currently considered as preliminary, this methodology will significantly improve over time.

*Correcting for Population Stratification in Case-Control Studies*

Population stratification, where cases are drawn from a different population than controls, continues to be a major issue in case-control studies, despite the enormous effort that is typically expended to adequately match cases and controls. Selection bias interferes with data interpretation by obscuring true associations and by generating false-positive associations that in reality are being driven by differences in the case/control populations. Genome-wide genotype data provide a straightforward solution to overcome this problem, as such data can be used to estimate principal component vectors that are then included as covariates in a linear regression model. This method effectively corrects for residual population stratification, and the incorporation of genome-wide data into standard epidemiological models will be an attractive tool for the future as single-nucleotide polymorphism genotyping costs continue to decrease.

*Environmental Risk Factor Study Design*

The area where genomics will have the most impact will be in environmental risk factor analysis study design. The myriad of variants that are reported to be associated with Crohn's disease has shown that genetics plays a far greater role in the pathogenesis of common diseases than previously thought [8]. Environmental factors do undoubtedly play an essential role in triggering disease and influencing phenotype, though the emphasis has now shifted to the concept of environmental agents working on a genetically susceptible individual [9]. Such gene-environment interaction will dominate the field of analytical epidemiology for the foreseeable future, though statistical techniques that adequately counter the enormous multiple testing involved in such studies remain to be resolved before the full power of this approach can be realized.

The classic risk factor study design of collecting as many environmental data as possible from as large a cohort as possible will give way to more tailored data acquisition based on knowledge of the underlying genetics and biology. For example, if it is known that variants within a particular biological pathway are responsible for causing a disease, then a parsimonious approach would be to focus data collection on environmental agents known to influence that pathway. Ideally, this targeted hypothesis approach will minimize the study costs by decreasing the sample size and by shortening the study time, while maximizing the chances of detecting relevant agents. A further intriguing possibility is prior selection of case and control subjects based on the presence or absence of a particular genetic marker with the specific aim of decreasing etiological heterogeneity, thereby increasing the ability to detect biologically relevant environmental risk factors. Epidemiological studies of Alzheimer's disease already stratify cohorts based on ApoE status, an approach that led to the identification of repetitive head trauma as a risk factor for developing dementia in carriers of the ApoE4 allele [10]. This approach will be greatly expanded upon in the design of future epidemiological studies.

*Genomics as an Epidemiological Tool*

Epidemiologists are already employing genetics as a tool of investigation, particularly in the area of infectious diseases. Sequencing of the genome of the severe acute respiratory syndrome virus was instrumental in tracing its phylogenetic lineage [11, 12], and a combination of genomic and epidemiological information allowed Chinese officials to trace the genotypic variation of the viral transmission paths [13, 14]. Similar approaches are being employed to understand the evolutionary biology and spread of bird flu and human influenza [15], both with potentially huge public health impact across the globe. As sequencing costs continue to decrease and whole-genome sequencing becomes a reality, genetics will be increasingly incorporated into neuroepidemiological studies.

*Epidemiology as a Genomics Tool*

Of course genome-wide association studies are not without their own problems, such as confounding arising from population stratification, the need for large sample sizes to detect minor effect alleles and inflated false-positive association rates arising from the several thousand tests that are an integral part of any such study [16, 17]. Epidemiologists can help geneticists overcome these problems, particularly by providing the infrastructure to collect large, well-phenotyped samples from affected and unaffected individuals drawn from similar ethnic backgrounds. Typically these cohorts are derived from population-based, natural history studies of particular diseases, often established many years ago prior to the development of the technology that underpins the genomics revolution. Indeed, there are already examples of how such projects have morphed into genomics in an effort to understand how genetic variation influences population susceptibility to disease. A genome-wide association study based on volumetric brain MRI and cognitive testing of 705 stroke- and dementia-free Framingham Heart Study participants identified significant correlation between SORL1 variants and abstract reasoning, and between CDH4 variants and brain volume [18]. Thus, it is true to say that neuroepidemiologists have long recognized the value of genomics in research, and have invested considerable resources to collect endophenotype data and to bank biological samples from population-based studies in the expectation of technological advances [3]. The future will see a tremendous return on their investment in this crucial infrastructure.

Determining the genetic variants that underlie complex diseases represents only the beginning, and 'translating' these discoveries to everyday clinical practice, as diagnostic tools and as therapy, will rely on carefully conducted, population-based epidemiological studies. The aim of these studies will be to understand the relevance of genetic variants associated with a disease within a population to disease within an individual patient. How many risk variants does an individual require before they are destined to develop a neurological disease? Do the variants merely affect age of onset, or do they also influence disease severity and outcome? How do these variants interact with each other to determine an individual's risk of disease, and what is the biological basis for this interaction? In complex diseases arising from multiple different loci in each individual patient, is changing the expression of a single variant sufficient to prevent disease in that individual? Is it too late to institute such an intervention at the time of first presentation, or should we undertake population screening and presymptomatic intervention? All of these questions must be considered before the advantages of our knowledge about genetics can take full effect. Longitudinal, prospective epidemiological studies are the ideal tool to address these issues in a meaningful, scientifically rigorous manner. An example of such a study is underway at the National Institutes of Health, where patients with Parkinson's disease due to mutations in the LRRK2 gene, identified as a key cause of familial and sporadic Parkinson's disease [19, 20], will be followed over a ten-year period to elucidate how symptoms develop over time (www.clinicaltrials.gov, NCT00467090). Such studies are likely to become commonplace in the future, as the genomic architecture of diseases is uncovered.

## Conclusion

In summary, there is a long-standing symbiotic relationship between epidemiology and genetics, which the current explosion in genomics will enhance by facilitating a more focused evaluation of environmental triggers, and which epidemiology will feed by providing well-phenotyped clinical samples. The result will be faster, cheaper and better tools for determining disease pathogenesis. The era of genomic epidemiology is truly upon us.

## Acknowledgements

This work was supported entirely by the Intramural Research Program of the NIH, the National Institute on Aging (project Z01 AG000949-02) and the National Institute of Neurological Disorders and Stroke.

## References


1 Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, San Giovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: Complement factor h polymorphism in age-related macular degeneration. Science 2005;308:385–389.

2 Hindorff LA, Junkins HA, Mehta JP, Manolio TA: A catalog of published genome-wide association studies. www.genome.gov/gwastudies (accessed May 26, 2009).

3 Khoury MJ, Millikan R, Little J, Gwinn M: The emergence of epidemiology in the genomics age. Int J Epidemiol 2004;33:936–944.

4 Frayling TM, Colhoun H, Florez JC: A genetic link between type 2 diabetes and prostate cancer. Diabetologia 2008;51:1757–1760.

5 Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Blondal T, Stacey SN, Helgason A, Gunnarsdottir S, Olafsdottir A, Kristinsson KT, Birgisdottir B, Ghosh S, Thorlacius S, Magnusdottir D, Stefansdottir G, Kristjansson K, Bagger Y, Wilensky RL, Reilly MP, Morris AD, Kimber CH, Adeyemo A, Chen Y, Zhou J, So WY, Tong PC, Ng MC, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Tres A, Fuertes F, Ruiz-Echarri M, Asin L, Saez B, van Boven E, Klaver S, Swinkels DW, Aben KK, Graif T, Cashy J, Suarez BK, van Vierssen TO, Frigge ML, Ober C, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Palmer CN, Rotimi C, Chan JC, Pedersen O, Sigurdsson G, Benediktsson R, Jonsson E, Einarsson GV, Mayordomo JI, Catalona WJ, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K: Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 2007;39:977–983.

6 MacReady N: New methods of gene analysis find genetic links among apparently unrelated conditions. Neurology Today 2008;8:36.

7 Lesnick TG, Sorenson EJ, Ahlskog JE, Henley JR, Shehadeh L, Papapetropoulos S, Maraganore DM: Beyond Parkinson disease: amyotrophic lateral sclerosis and the axon guidance pathway. PLoS ONE 2008;3:e1449.

8 Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–962.

9 Perera FP: Environment and cancer: who are susceptible? Science 1997;278:1068–1073.

10 Nicoll JA, Roberts GW, Graham DI: Apolipoprotein E epsilon 4 allele is associated with deposition of amyloid beta-protein following head injury. Nat Med 1995;1:135–137.

11 Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL: The genome sequence of the SARS-associated coronavirus. Science 2003;300:1399–1404.

12 Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ: Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 2003;300:1394–1399.

13 Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, Su ST, Chia JM, Ng P, Chiu KP, Lim L, Zhang T, Peng CK, Lin EO, Lee NM, Yee SL, Ng LF, Chee RE, Stanton LW, Long PM, Liu ET: Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 2003;361:1779–1785.

14 Zhao GP: SARS molecular epidemiology: a Chinese fairy tale of controlling an emerging zoonotic disease in the genomics era. Phil Trans R Soc Lond B Biol Sci 2007;362:1063–1081.

15 Nelson MI, Holmes EC: The evolution of epidemic influenza. Nat Rev Genet 2007;8:196–205.

16 Simon-Sanchez J, Singleton A: Genome-wide association studies in neurological disorders. Lancet Neurol 2008;7:1067–1072.

17 Schymick JC, Talbot K, Traynor BJ: Genetics of sporadic amyotrophic lateral sclerosis. Hum Mol Genet 2007;16:R233–R242.

18 Seshadri S, DeStefano AL, Au R, Massaro JM, Beiser AS, Kelly-Hayes M, Kase CS, D'Agostino RB Sr, Decarli C, Atwood LD, Wolf PA: Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham study. BMC Med Genet 2007;8:S15.

19 Gilks WP, Abou-Sleiman PM, Gandhi S, Jain S, Singleton A, Lees AJ, Shaw K, Bhatia KP, Bonifati V, Quinn NP, Lynch J, Healy DG, Holton JL, Revesz T, Wood NW: A common LRRK2 mutation in idiopathic Parkinson's disease. Lancet 2005;365:415–416.

20 Paisan-Ruiz C, Jain S, Evans EW, Gilks WP, Simon J, van der BM, Lopez DM, Aparicio S, Gil AM, Khan N, Johnson J, Martinez JR, Nicholl D, Carrera IM, Pena AS, de Silva R, Lees A, Marti-Masso JF, Perez-Tur J, Wood NW, Singleton AB: Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron 2004;44:595–600.