Roses, J Pharmacogenom Pharmacoproteomics 2012, 3:3
DOI: 10.4172/2153-0645.1000e120

Journal of Pharmacogenomics & Pharmacoproteomics

Editorial Open Access

Post-GWAS: Phylogenetic Analysis in the Hunt for Complex Disease-Associated Loci

Allen D Roses*

[1]Jefferson-Pilot Professor of Neurobiology and Neurology, Duke University, Durham, NC, USA
[2]President and CEO, Zinfandel Pharmaceuticals, Inc, Chapel Hill, Durham, NC, USA

What constitutes reliable disease-relevant genetic association data? Scientific research arena has gone through remarkable transformation over the past decade in addressing this conundrum. It was recognized that more than 95% of reported disease associations in the pre-Genome Project years were neither consistently validated nor qualified in follow up investigation beyond the initial publication. In response, editors of the major scientific journals required biological or clinical data to substantiate the statistical findings.

Subsequent to the explosion of Genome Wide Association Study (GWAS) data, yet a different scenario has emerged, reversing the standards to which the community had become accustomed. Multiple, barely significant SNP associations are routinely reported and widely accepted as new markers associated with complex diseases without much mechanistic or biological support. Most of these are based on agnostic statistical GWAS results, followed by a confirmatory evaluation of the top SNPs or SNPs in nearby genes. The elevation of small effect SNP data has led to premature claims for the effect of these genes on the pathogenesis of disease. Furthermore, the community has fallen into a pattern of circular investigations, where "missing heritability" is to discovered by larger GWAS studies to identify even more small effect size SNPs. The statistical conclusions race far ahead of the biology, or even more comprehensive methods of genetic characterization and statistical analysis.

Alzheimer's disease (AD) provides perhaps the most instructive case study. More than a dozen GWAS studies had confirmed the extraordinary statistical significance of the "*APOE*" linkage disequilibrium (LD) region in association with risk and age of AD onset [reported with as high as $p < 10^{-157}$ in one large study [1]. However, the clinical utility of the *APOE* e4 allele, shown to be associated with younger age of disease onset and with increased disease risk that is dependent on allele dose, is limited. The *APOE*ε4/4 genotype, which provides an approximately 14-fold increase in disease risk, is carried by only 2% of the Caucasian population. Carriage of *APOE* ε4 is estimated to account for approximately 20% of disease heritability. This led investigators to look across the rest of the genome for other gene variants that are statistically associated with AD, rather than seek explanation to the enormous and unparalleled statistical support for the "*APOE*" LD region. Most of the non-*APOE* genomic signals resulting from GWAS have barely reached statistical significance after correction for the number of tests and, more concerning, there has been precious little translation to biological or medical experiments to support their qualification as a genetic test.

It has long been demonstrated that genetic variants in strong LD with each other can reflect the association carried by the entire haploblock, whether or not a biological significance for each of these variants exists. Recently, researchers in our group used phylogenetic analysis to further elucidate the nature of the association signal around the *APOE* LD region and pinpointed a small stretch of sequence that segregated with disease risk [2]. AD patients were concentrated in the clade where cognitively normal controls were infrequent. That is, that AD cases clustered in clades because of the shared evolutionary history of specific, disease-associated, haplotypes. Phylogenetic analysis has rarely been applied to gene hunting for human disease, but has been in use for almost three decades as a way of rapidly recognizing pathological changes in the genetic sequence of rapidly mutating organisms like viruses. This methodology is in fact standard for defining the emergence of new strains of influenza viruses or HIV. As large-scale DNA sequencing becomes a more affordable commodity, phylogenetic analysis should become widely employed for discovering disease-associated variants, including structural variants, which are related to a phenotypic variable that can be accurately measured - like age of onset.

In the case of AD, one clue that led us to our discovery of another disease-associated variant in the *APOE* region was that the three-allele two-SNP *APOE* system could not mathematically account for the p values coming out of GWAS publications. It should be noted that the two SNPs that determine *APOE* genotype are not included on the GWAS platforms [so as to avoid the misuse of *APOE* alleles as diagnostics]. Even in the absence of direct *APOE* measurements, all of the GWAS papers [ours was the exception] attributed the signal in this region to *APOE*. We reported the association of a long, polymorphic poly-T repeat , rs10524523, in an intron of *TOMM40* with age of AD onset. *TOMM40* encodesthe pore subunit of the translocase of the outer mitochondrial membrane complex, a protein that is essential to mitochondrial function. The poly-T polymorphism stratifies the age of AD onset distributions for 97% of the Caucasian population (all *APOE* genotypes except for *APOE*ε2/4] and *APOE*ε2/2, rather than the 2% of this population that possess the *APOE*ε4/4 genotype]. While some studies have refuted the association of *TOMM40* with age of AD onset, these studies were conducted in collections of AD patients with inconsistently ascertained age of onset data. The pivotal studies are currently underway, using prospectively observed age of onset data that is rigorously ascertained in longitudinal epidemiological cohorts, as well as in prospectively followed clinical series.

The trial is designed to achieve regulatory qualification of rs10524523 genotype as a biomarker to predict high or low risk of

***Corresponding author:** Allen D Roses, Jefferson-Pilot Professor of Neurobiology and Neurology, Duke University, Durham, NC, USA, Tel: 919 660 8065; Fax: 919 681 9289; E-mail: allen.roses@duke.edu

Received April 25, 2012; **Accepted** April 26, 2012; **Published** April 28, 2012

Citation: Roses AD (2012) Post-GWAS: Phylogenetic Analysis in the Hunt for Complex Disease-Associated Loci. J Pharmacogenom Pharmacoproteomics 3:e120. doi:10.4172/2153-0645.1000e120

developing mild cognitive impairment (MCI) in normal individuals within the next 5-7 years in the relevant age group being studied, and will occur simultaneously with the evaluation of a drug to prevent MCI of the AD type. First subjects into this population-based prevention trial, with subjects stratified by age 67-83, rs10524523 genotype and the presence or absence of *APOE2*, is planned for the third quarter 2012 after US Food and Drug Administration review of the Investigational New Drug application. This clinical trial is sponsored by an alliance of Zinfandel and Takeda Pharmaceutical companies and is based on a primary prevention design. The prognostic biomarker will not be used clinically until completion of the double-blind, pharmacogenetic-assisted clinical trial.

Phylogenetic mapping, based on highly accurate, phased DNA sequence data that includes structural polymorphisms (indels, homopolymers, etc.), has enormous potential for identifying genetic variants that are important in the development or progress of other common diseases. However, it is still, unfortunately, in the "ignore" stage. Publishing and grant funding is biased toward the conventional, and insufficient, GWAS approach.

The time will come when "this too will pass." There are examples of using phylogenetic methods to identify the specific genetic variants directly related to phenotypic variation. However, only more recently has phylogenetic analysis of DNA sequences from a single species been recognized as a powerful approach for genotype/phenotype association studies. In fact, in areas of low, or no, recombination (for example, the *APOE-TOMM40* region) use of phylogenetic methods is likely a much more powerful approach to finding phenotype-related variants than single SNP, or even standard haplotype, methods. Not only does phylogenetics greatly reduce the number of statistical association tests relative to single polymorphic site analysis, and thus increases statistical power, but the methodology is more likely to identify the set of mutations, or evolutionarily-related haplotypes, that could be causative. As Templeton states, for regions of LD, associations identified by SNP by SNP analyses may even be misleading [3].

For human disease studies, it is important to use carefully documented disease-phenotype variability data, rather than simple disease diagnostic labels that mask the underlying heterogeneity.

Recognizing the role of accurate disease phenotypic data [and not just series of disease-labeled patients with uncertain diagnostic informaton] seems to have worked. To have travelled from the initial discovery of a disease-associated genetic marker to translation to a pharmacogenetic-assisted Phase III, primary prevention trial, within three years and under the academic radar, provides an example of the true power of genetic biomarkers in the field of translational medicine. Let's hope that it does not take the full duration of the primary prevention study for the community to realize and accept the point that post-GWAS is not just chasing the next high-throughput sequencing technology or accruing ever larger, shallowly-phenotyped cohorts . It is rather a program of hypothesis-driven analyses grounded in accurate clinical data and reliable sequence information - including the full gamut of structural variants, in the example presented here a variable-lengthpoly-T homopolymer, which may not be accurately accounted for by the currently-available, "next-generation" sequencing methods.

References

1. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, et al. (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 41: 1088-1093.
2. Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, et al. (2010) A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J 10: 375-384.
3. Templeton A (2010) The Diverse Applications of Cladistic Analysis of Molecular Evolution, with Special Reference to Nested Clade Analysis. Int J Mol Sci 11: 124-139.