Reviewer's report

Title: A Centralized Open Access Database of Genome-wide Association Results

Version: 1 Date: 3 September 2008

Reviewer: Scott Saccone

Reviewer's report:

SUMMARY

The authors present a well-written and informative account of a new resource devoted to the amalgamation of results from genome-wide association studies (GWASs). They have carefully constructed an annotated database derived from a large number of GWASs. They report a number of novel findings that arise from a meta-analytic approach, and suggest additional protocols for reporting GWAS results that would enhance this approach. However, there are certain aspects of the methods that require modification in order to be analytically sound, and some of the results are difficult to interpret.

MAJOR COMPULSORY REVISIONS

1) When reporting the “enrichment” of signals, whether this is with respect to physical regions or logical sets of genes such as pathways, the number of signals from a given sample must be corrected for linkage disequilibrium as estimated in the appropriate HapMap sample. Otherwise there will be bias towards regions of strong LD.

MINOR ESSENTIAL REVISIONS

1) The statistical test for over-representation of biological pathways must be presented with greater detail, so that the interpretation is clear. What does \( p < 4.6 \times 10^{-14} \) for cell adhesion functions really mean? Does it mean that this set of 650 genes contains one or more true associations? The authors should explain if this method is biased by the size of the gene set.

2) The authors should explain how the filter 0.001 was selected in the second paragraph of page 4. If this filter is used for the public version of the database, they should explain why users should not have the freedom to filter using different (higher) thresholds.

3) In the section “Informatics challenges…”, the authors claim it is problematic that different studies use annotation from different sources, such as different builds of dbSNP. When creating this combined GWAS database, the authors should explain why we should not simply ignore the original bioinformatic annotation and use the most current annotation available, and use only rs ID and p-value as the primary inputs.
4) The rationale for reporting the fraction of “associated SNPs” within genes should be justified. Given the results, is the implication that there should be a shift in the initial design of a GWAS toward genic regions, or that genic SNPs should have greater priority when selecting SNPs for further study after a GWAS?

5) On page 9 in the first paragraph, perhaps it should read “5E-08 <= P <= 0.0001”. As written, it seems redundant. The same goes for the legend to figure 1.

6) Several column headings in the figures seem to lack a description. For example, the authors should explain what is meant by “Perfect CEU proxies on other platforms” in Table 3.

7) Privacy concerns should be more thoroughly addressed. It appears that association results, not genotype data, are made available. But it should be made clear, such as in the text at the bottom of page 16, whether allele frequency estimates, particularly those in case samples, will be made publicly available. If so, this could raise privacy concerns: see “Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays” by Homer et al., PLoS Genetics, Vol. 4, 2008.

DISCRETIONARY REVISIONS

1) A key question after an association study is “are we done yet?”. A useful addition to this database would be for each SNP, phenotype and ethnic population to show some measure of the likelihood of a true association (main effect) given the data. For example, if we have SNP in a gene we are very interested in, although it might have been tested already, perhaps there was insufficient power to detect an association with an odds ratio no greater than 1.2, and therefore further study could be warranted. We could then ask the same question about the entire gene: have all common SNPs in or near the gene been tested with sufficient power?

2) On page 14 – “…to weight SNPs with a priori evidence for association” – you may also want to consider the papers “Systematic biological prioritization after a GWAS…” by Saccone et al. (2008) in Bioinformatics, and “A pragmatic suggestion for dealing with results…” by Curtis et al. (2007) in BMC Genetics.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:

I, Scott Saccone, am listed as an inventor on a patent (US 20070258898) held by Perlegen Sciences, Inc., covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction.