Reviewer’s report

Title: Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information.

Version: 1 Date: 10 June 2009

Reviewer: Bruce Aronow

Reviewer’s report:

The manuscript “Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information by Yonqing Zhang et al” undertakes a systematic analysis of human genetic association data with mouse genetic analyses. The authors have developed the valuable human gene association database resource and in this manuscript describe some of the overarching themes that emerge with respect to categories of disease as defined by either the NLM MeSH classification or MGI MP ontology.

1. The authors principally examine gene set overlaps and define a distance function based on a normalized measure of set overlap. Based on these group-to-group distance measures, tree structures reflecting the matrix of group similarities have been constructed and are presented as dendrograms.

2. Using the Ward’s algorithm in SAS JMP, a series of hierarchical tree views have been generated. These are reasonable approaches, but the kinds of conclusions that a reader should draw from these are not as clear as one would hope to glean from these exercises and resulting resources. Where do we go from here? What is the actual genetic basis for different disease group differences? What does it mean that there are fractional overlaps and fractional differences?

3. If the authors could provide readers/users with some additional basis for thinking about this more concretely it would be helpful. I am also not sure about the relative distances: in Figure 2 a specific question here is what does it mean that there seems to be a closer relationship between hematopoietic, cardiovascular and metabolic categories than to immune and inflammatory categories? In my analysis, the gene overlaps are very high between immunoinflammatory diseases and hematopoietic but much less so with cardiovascular and metabolic.

4. The supplementary tables are very good and look quite useful, but their generation from the GAD should ideally be made fully automated and, like the JAX MGI, there should be a consistently structured report that is sortable or parsable with respect to the evidence for the association including if possible, a representation of the specific causal/associated allele.
5. The comparison and update status of GAD with OMIM is not as clear as one would like to see. The description of GAD in the Methods section is good, but perhaps this can be clarified a bit more. How complete, up-to-date, and filterable are the results from GWAS studies? Are there interesting leads or insights that can be developed from gene-disease associations that are in one, but not the other, or are these mostly reflective of asynchronous updating processes? The same is of course even more important in the comparison of GAD with mouse phenotypes, but before this can be convincingly presented, one wants to understand more clearly how complete the gene lists are for each human disease category associations?

6. From my analysis of the gene listings of Table 1 and Supplementary Table 1, looking for functional and annotation enrichments and overlaps (particularly for “Disease” and “Human Phenotype” using the ToppGene resource (http://toppgene.cchmc.org/), I was not sure how complete the provided tables of human disease group genes listings are? Are these tables restricted for any features or annotations within the GAD for human disease-associated genes? If there are differences between GAD-generated gene lists for positive gene associations and those of resources such as OMIM or Human Phenome, how should these be thought about? Why wouldn’t the GAD want to include both?

Additional questions:

7. In Table 1, why would the list of genes for Pathological Conditions, Signs, and Symptoms be fewer than the sum of gene lists in categories such as Cardiovascular Disease, Neoplasms? This is probably not critical but it should be explained how this happens.

8. The authors mention a number of interesting issues related to origins of similarities and differences of resulting disease characteristics and mouse phenotypes in humans versus mice. Some of these have been introduced previously in the original description of the GAD in Nature Genetics. Perhaps the authors could discuss this a bit further in light of their overall analysis. To what extent are the differences based on the occurrence of different kinds of alleles of given genes versus true differences in species specific gene functions?

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 declare that I have no competing interests