Reviewer's report

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

Version: 1 Date: 14 September 2009

Reviewer: Bing Zhang

Reviewer's report:

General

Zhang and colleagues described a new gene set database with human disease and mouse phenotype based gene sets. They also illustrated potential application of the database in comparing diseases/phenotypes based on associated genes. Overall, it is a useful resource. However, the manuscript is not well written. Serious revision is needed to make this manuscript more concise and easier to follow.

Major Compulsory Revisions:

1. The “Summaries of genes and phenotypes in human and mouse” section is very lengthy. This section can be summarized in a few paragraphs describing the statistics (e.g. Tables 1 and 2) in the four tables (Human gene to disease table, Mouse gene to phenotype table, Human disease to gene table, and Mouse phenotype to gene table) and major findings from individual tables. Paragraph 1 under Results belongs to disease to gene table part, and paragraph 2 under Results belongs to gene to disease table part. Tables 4, 5, 7, 8 are not necessary, corresponding supplemental tables should be enough. URLs to the interactive version of the tables should be provided in the text instead of a simple link labeled “here”.

2. In the “comparisons and analysis using disease and gene lists” section, it is not clear why the authors chose to use both the phylogenetic tree analysis and the hierarchical clustering analysis. My understanding is that they both support the same conclusion that gene set similarity analysis can group similar disease/phenotypes together. If the two analyses complement each other and reveals distinct patterns, a discussion is needed. Otherwise, only one method is necessary for the manuscript. The section title mentioned “analysis using disease gene lists”, however, there is only a brief description at the end of the section on GSA analysis. This analysis part needs to be expanded or removed.

3. One potential and attractive application of the database is to cluster both human disease gene sets and mouse phenotype gene sets together to explore similarity between human diseases and mouse phenotypes. The authors mentioned in the abstract that “…human disease as compared to itself and in the context of mouse genetic models of disease”. This analysis is necessary to support the conclusion.
Minor Essential Revisions:

1. Page 7, the notations for the distance formula are not clear. The min function doesn’t look right.

2. Page 7, the formula for the Fitch analysis needs to be described.

3. Page 8, the distance formula for the Ward’s method needs to be better described.

4. Page 10, paragraph 2, “asthma” seems like a child term of “immediate hypersensitivity”, “macular degeneration” and “choroidal neovascularization” don’t seem to have direct parent-child relationship. It will be helpful to provide the MeSH tree number to show their relationship.

5. Page 11, line 8, what is MP:#####?

6. It is difficult to follow the numbers of the gene sets. For human, 1318 sets on page 13 but 480 sets on page 16 and 19. For mouse, 5143 sets on page 14, 1056 on page 17, and 2067 on page 19. How were the gene sets selected?

7. Page 15, paragraph 3 fits the Discussion section better, no results to support most of the suggested applications.

8. Page 21 the first paragraph, the wording is very confusing, e.g. what is “in broad based comparative analysis utilizing network and approaches”? Which network?

9. The section title “comparisons and analysis using disease and gene lists” is not clear either. Should it be “using disease gene lists”?

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**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.