Author's response to reviews

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

Authors:

Yongqing Zhang (zhangyon@grc.nia.nih.gov)
Supriyo De (desu@grc.nia.nih.gov)
John R. Garner (garnerir@nia.nih.gov)
Kirstin Smith (keersteen@yahoo.com)
S. Alex Wang (wangal@mail.nih.gov)
Kevin G. Becker (beckerk@grc.nia.nih.gov)

Version: 5 Date: 8 January 2010

Author's response to reviews: see over
January 8, 2010

Editor, BMC Medical Genomics:

We would like to resubmit the revised manuscript Zhang et al. “Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information” for publication as an article in BMC Medical Genomics.

We have made changes to tables and figures as suggested by the core-pre accept editor.

This research was supported by the Intramural Research Program of the NIH, National Institute on Aging, and the NIH Center for Information Technology, National Institutes of Health.

Sincerely,

Kevin G Becker

Kevin G. Becker Ph.D.
Suite 100, Room 4B132
251 Bayview Boulevard
Biomedical Research Center
National Institute on Aging
National Institutes of Health
Baltimore, MD 21224
beckerk@grc.nia.nih.gov
410-558-8360
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?

The dendrograms and clustergrams provide evidence in both the human and mouse case that the underlying assignments to phenotypes have disease relevance and informatic value. This is not currently accepted, at least on the human side. As we suggest in the paper this should give greater confidence to algorithm developers, web tool developers, etc. to use these files in creative and robust ways. This may include annotation tools, gene set analysis tools etc.

The actual genetic basis for group and fractional comparisons will take further analysis than what we have done here. One could do an entire paper on the differences and overlap between any two major groups (which we have considered).
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?

It may be that the hematopoietic MeSH annotation reflects classical blood and red blood cell disorders (sickle cell anemia, porphyrias, etc.) while any disorders that include white cells, inflammatory cells, etc., were tagged in MeSH with immune and inflammatory annotation.

In my analysis, the gene overlaps are very high between immunoinflammatory diseases and hematopoietic but much less so with cardiovascular and metabolic.

It is of course likely that the relative positions of disease groups would move around in space depending what factors or values are included in the analysis or weighted in the comparisons. Possibly if different ontologies, such as SNOMED were used, different groupings would appear. If average relative risk or ODDS ratio were used in some way, the groupings would change. We are not sure about the details of your analysis and cannot comment on it at this time.

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.

Although this is a laudable goal this is beyond the scope of this work. Automated reporting and comprehensive detailed allele reporting would require funding which is way beyond our means. Given that we have no direct funding for this project and have never had funding.

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. We have clarified differences between OMIM and GAD P4 L3-5.

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?

GAD is not necessarily complete and makes no representation that it is. We work essentially as does the genome project from frozen dated builds.

The question of completeness also begs the question of what complete is? As GWAS studies move forward, two things are clear. More rare alleles of genes are being found, and effect sizes are going down. Therefore, in my view it is unlikely, on average, that future data will change the overall composition of the groupings. Individual new genes may be dramatic. Different algorithms will reveal different groupings or relationships. But in my view,
whether if the database is ~80% complete or ~100% complete in the aggregate would not dramatically alter the major dendrograms. Also, it is our hope that when we get this paper out of the way, we will embark on an updating effort in GAD.

OMIM is a very different database in scope and style than GAD. The majority of records in OMIM archive rare mendelian disorders while GAD records common complex diseases. These are two very different classes of diseases. Although more recently OMIM has included common diseases, it reflects the statistical approaches and large effect sizes of classical genetics. In our view we need newer models which accommodate low effect sizes.

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:

The data provided here from GAD is not restricted for any parameters other than positive association and MeSH annotation. As mentioned above in #5, GAD and OMIM are very different datasets and archive very different types of records in different ways.

I cannot find any reference or mention of Human Phenome, other than a paper in 2003 by Nelson Freimer & Chiara Sabatti suggesting a joint effort.

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.

I am not exactly sure. It may be due to incomplete or inconsistent assignment of MeSH terms in the MeSH ontology or possibly due to the directed acyclic nature of MeSH.

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?

In our view the major differences between the human and mouse data collected here are related to the descriptive qualities of the ontology terms used in each case. In the human case MeSH terms typically describe aggregate end stage clinical phenotypes while in the mouse, the mammalian phenotype ontology used specify multiple observed phenotypes of a resulting mutation. For example, in human you might get multiple sclerosis, while in mouse you typically might get subphenotypes or endophenotypes or altered membrane potential or unsteady gait, or altered T cell ratios. In addition, the mouse genetic models are skewed to gene knockouts with more profound phenotypes that typically present earlier than end stage human diseases.
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.

   We feel that the audience for this paper is as much everyday biomedical researchers or possibly even clinicians as much as computational biologists. Large text files with thousands of rows can be intimidating for the uninitiated. Since this is not primarily a computational journal we feel that tables 4, 5, 6, 7, 8 are necessary to introduce the material to non-computational biologists. In addition, in a printed pdf of this paper the supp tables would not typically print out. Therefore, the reader would not have any point of reference for major sections of the content while reading.

   URLs to the interactive version of the tables should be provided in the text instead of a simple link labeled “here”.

   Corrected Table designations are provided in the text, and URL links have been added at the end in the list of supplemental tables.

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.

   After careful consideration, we agree that the hierarchical cluster is redundant to the dendrogram. We have shortened the discussion of the HC and have moved all the figures to supplemental figures.

   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.

Our parallel paper is referenced which describes the use of these gene sets in gene set analysis of microarray data. We have moved this to the discussion and expanded the portion of the text (p21, L13-12) to better describe this use. We can provide the manuscript of the other paper to the reviewer if necessary.

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.

While it is seemingly logical and straightforward we couldn’t get a good outcome the way we have tried. This may be due to different levels at which the mouse and human are annotated (see response #8 above).

It is likely that once published, clever computational biologists will come up with creative ways to integrate the mouse and human tables.

However, we did have considerable success integrating mouse with human and human with mouse when used in gene set analysis of microarray data. This is described in a parallel paper on Gene Set analysis. We can provide a copy of that manuscript if the reviewer wishes.
Minor Essential Revisions:

1. Page 7, the notations for the distance formula are not clear. The min function doesn't look right.
   This has been corrected in the manuscript.

2. Page 7, the formula for the Fitch analysis needs to be described.
   The formula has been replaced by a simplified version from the PHYLIP manual.

3. Page 8, the distance formula for the Ward's method needs to be better described.
   Description for the Ward's method has been added.

4. Page 10, paragraph 2, “asthma” seems like a child term of “immediate hypersensitivity”.
   The terms Asthma and Immediate hypersensitivity are in different branches of the Disease branch in MeSH. However, clinically, immediate hypersensitivity is a clinical subphenotype of asthma.

   “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.

   Macular Degeneration and choroidal neovascularization do not have a parent child relationship in MeSH. Regardless of the MeSH designations, choroidal neovascularization is a clinical subphenotype of macular degeneration. Our gene sharing algorithm highlights that relationship.

5. Page 11, line 8, what is MP:#####? It is the format for the mammalian phenotype codes from the JAX labs phenotype tables, used in the mouse phenotype gene set tables.
   We have changed it to 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.
   As mentioned on page 13, in this summary there are 1317 human gene sets total.
   There was a minimum cutoff of 3 genes each in each gene set to build the dendrogram for human, a minimum of 10 genes per gene set for mouse..

   For mouse, 5143 sets on page 14, 1056 on page 17, and 2067 on page 19. How were the gene sets selected?

   There were 5142 total mouse gene sets. Ten genes per gene set were used as a cutoff in the mouse case, since a cutoff of 3 would produce a computational load that would take an
enormous amount of computational time, (months). Hierarchical clustering was more computationally efficient and 2067 gene sets were used, using a cutoff of 3 gene sets.

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

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?
   Agreed, network was changed to the more general “systems”.

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