Author's response to reviews

Title: Genotator: A disease-agnostic tool for genetic annotation of disease

Authors:

Dennis P Wall PhD (dpwall@hms.harvard.edu)
Rimma Pivovarov (rimma_pivovarov@hms.harvard.edu)
Mark Tong PhD (mark_tong@hms.harvard.edu)
Jae-Yoon Jung PhD (jj77@hms.harvard.edu)
Vincent A Fusaro PhD (vfusaro@hms.harvard.edu)
Todd F DeLuca (todd_deluca@hms.harvard.edu)
Peter J Tonellato PhD (peter_tonellato@hms.harvard.edu)

Version: 2 Date: 16 September 2010

Author's response to reviews: see over
September 16, 2010

Dear Editors,

We would like to thank you and the reviewers for the recent thoughtful review of our manuscript entitled “Genotator: a disease-agnostic tool for genetic annotation of disease.” We have carefully revised the manuscript in response to the reviewers’ critiques and provide a point-by-point summary of our changes in the pages below. As requested, we have paid particular attention to points 3 by reviewer 1 and 7 by reviewer 2, i.e. the need for further validation of the results, adding a 3rd use case (Alzheimer Disease) as a consequence. We have also significantly improved the speed and usability of the web site, and have elected to make the source code available.

We hope will find our revised manuscript an improvement over the previous and now suitable for publication in its present form. Thank you very much for your consideration.

Sincerely,

Dennis P. Wall

Yours truly,

Dennis Wall, PhD
Assistant Professor
Department of Pediatrics &
Director, Computational Biology Initiative
Center for Biomedical Informatics
Harvard Medical School
As requested, we have reformatted the abstract and have added a competing interest statement. Below please find our point-by-point responses to the reviewers’ comments. We found the reviews extremely helpful and have carefully revised our manuscript to address the concerns.

Responses to Reviewers (reviewers comments in bold)

Reviewer 1

The scoring formula relies heavily on Yes/No labels from GAD. To my knowledge, GAD only added a few records since 2008, meaning that GAD has not updated their data.

How do you address this problem?

GAD recently updated their database (August 1, 2010). This update has increased the number of records 2 fold from 40,000 to over 84,000 records. Because Genotator integrates this information dynamically, it will remain up-to-date with GAD and the other databases, ensuring consistency. While the scoring algorithm does utilize GAD Yes/No labels, the weight of these labels on the score will depend on the availability and number of association studies. In many cases, the availability and number is low. For example, a search for “autism” returns 500 entries, less than half of which have Y/N labels. In the absence of such data, the weight will shift to the other parameters in the score, a property of the scoring system that we considered desirable, as it will provide some ability to prioritize even in the absence of further granularity (and in these cases the prioritization will be well correlated with the valuable gene prospector score).

On a more general note, we intended the Genotator score to be a simple guide rather than a definitive ranking tool. We provide numerous attributes in the Genotator disease reports not only to enrich the information content for the users but also to enable the users to rescore on the basis of alternative formulas (and to entirely disregard the Genotator score). We have added language to emphasize this point in the manuscript.

I tried a couple of examples using both Genotator and Gene Prospector. It seems that the returning results were quite similar. What is the added value of Genotator over Gene Prospector in terms of prioritization of the candidate genes?

We incorporated the Gene Prospector Score into the Genotator scoring algorithm because of its inherent value, and thus expect that in many cases the two scores will be correlated. However, the other parameters of the score, including the GAD Y/N parameters, will alter the rank order. This will become increasingly more common as the number of association/replication studies increase. This serves to add another level of granularity that can be used to choose among a set of highly ranked genes, e.g. for experimental validation. In addition, because Genotator agglomerates data from 11
total resources, the gene lists and the rankings will deviate from Gene Prospector, especially toward the middle to lower end of the candidate gene rankings.

Since the parameters constants in the formula were obtained based on the two test cases (autism and Parkinson). I am afraid they might cause over fitting issue. Try a different case (for example, Alzheimer and compared with Top gene list from Alzgene (http://www.alzgene.org/)) with the same set of the parameters and see how the tool performs.

The constants were selected based on a series of tests including several other diseases. This empirical assessment was most useful for deciding on the number of references; we wanted to avoid introducing ascertainment bias while still boosting genes replicated in multiple studies. We have improved our explanation of this in the manuscript.

In addition, we agree that a third use case would be valuable to demonstrate the utility of the tool and sensitivity to the parameter settings. Following your sound advice directly, we have conducted a third study with the AlzGene database. The results were even stronger than in the other two use cases, with 90% concordance among the top 10% percentile of Genotator. Taken together, we have solid evidence that Genotator provides rankings similar to that available from heavily annotated sources, without risk of over-fitting.

Reviewer 2

I think that this paper makes a useful contribution in the area of assembling a list of genes that are involved in a disease, based on aggregating existing databases. There could be a more complete literature review of other methods for doing this, including algorithmic ways (iHOP and others) as well as natural language text processing for extracting disease-gene interactions. This could be a paragraph, but would give the reader a sense of other options for getting candidate gene lists.

We agree with this comment and have altered the Discussion section to include references to alternative strategies for linking genes to disease. We now reference iHop, as well as text-mining strategies that have been developed for automatic extraction of gene-disease association from the literature.

The website was beta and was very slow, but it eventually worked. I assume it may be made faster if they move it to a production machine. This is not just the email that was slow, but even rendering the home page, and switching tabs on that page, so it is really at the border of usability.

We have taken steps to speed the web interaction, including porting all of Genotator to a production machine. The results generation is dependent on the size of the results
returned from each database as well as the response times of those databases. Thus, we elected to use an asynchronous email process to improve user experience.

The website is available in Beta right now. I do not seem that the code is available, and I’m not sure if that is what you require. I would support making the code available, but right now it seems that only the website functionality is available. Particularly given the slow response speed of the site, it would be important to give potential power users the option of running it locally. A simple free/cheap license for research use seems reasonable.

We elected to build a web resource rather than release the code principally because the code interacts with a host of external websites and must be adapted to unpredictable changes in the formats of those external sites. By centralizing via the webtool, we can monitor the resource for failures, fix them rapidly, and thereby serve a larger number of users faster and more consistently. In addition, one of our objectives was to make Genotator a growing repository of genetic data for many human diseases, and a resource that grows and updates as the community uses it. By directing users to our site, the repository will continue to keep-pace with the rate of research being done on various human diseases and therefore continue to grow in value. By centralizing in this way, we can help prevent duplication of effort, ensure faster delivery of information, increase data sharing, and ultimately facilitate large cross-disease studies.

We have also included a batch submission option with Genotator that will allow power users a user to submit numerous disease queries simultaneously. These jobs are submitted in batch to our research cluster so that the response time will always be as long as the time to run an individual disease query.

In addition, we have included a note in the manuscript that the source code can be provided “as is” upon request.

Finally, as indicated above, we have moved Genotator to a significantly faster machine and have optimized queries so that the site is significantly more responsive than the previous development version.

The validation is limited—two diseases and then comparing the lists to some unclear gold standard. Actually, the paper should state more clearly what the gold standard is, why it is a gold standard, and perhaps provide quantitative measures of the agreement of their candidate gene list with others. A third example would be nice, but perhaps not mandatory.

We have stated more clearly why SFARI gene and PDgene are good baselines for comparison. The main reason is that they are updated regularly and monitored by experts in the field. The SFARI gene has manual curators who monitor the literature carefully and only include genes that have strong positive support. The PDgene database
uses a three-step inclusion process that also involves human supervision. Although these resources are not truly gold standards, they represent the standard we wanted Genotator to achieve. That is, we wanted to optimize Genotator to match carefully monitored and trusted gene databases, a strategy very similar to that used in the GeneProspector paper (Yu et al. 2008) in which they evaluate the performance of their system using Parkinson as a query and the PDGene database. In their evaluation they report that GeneProspector contains 2 of the 13 Top Genes listed by PDGene at that time. Gold standards, especially for multigenic diseases whose candidate genes have not been fully enumerated, are either not available or too incomplete.

Following your excellent advice (and the advice of reviewer 1), we have also elected to add a third use case, Alzheimer Disease, to further demonstrate the utility of Genotator. Using a similarly high quality and manually curated resource, AlzGene, we found an even stronger correspondence, with 90% agreement among the top 10% percentile of Genotator, and 75% similarity overall.

It says “Java, Python” and so it could be a little more explicit about how the code runs, what hardware is required. Also, is there a plan for maintenance—the abstract says “real time” which implies that they are scraping data from these sites—what will happen if they change their formats, etc…? Some discussion of this might be reasonable.

We have added additional details to the algorithm section (under Implementation). We include details on the code design including the distribution of jobs across a research cluster to enable batch processing for power users.

I don’t know about the word “comprehensive.” I would think that comprehensive would include primary NLP analysis of text as well as the analysis of uncurated databases like GEO to pull out candidate genes.

Agreed. We have elected to remove it from the title.

There are a few too many “due to” for my tastes, but the writing is clear. We kept the “due to” usage to a minimum and also cleaned up the language throughout the manuscript.