Reviewer's report

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

Version: 1 Date: 14 August 2010

Reviewer: Russ Altman

Reviewer's report:

1. Does the software address a novel task? Alternatively, if there is already software available that performs this task, does the software outperform it in terms of speed, reliability, efficiency, or breadth of application?

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.

2. Is it easy to use?

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.

3. Does it satisfactorily address the task or application the authors intend?

I think it is reasonable. There is a large reliance on a few databases which begs the question of whether or not this really adds a lot. The real added value, I suppose, is that they check the other 9 and make it easier for a user to do that.

4. Is the software freely available for non-commercial use (note that this is a condition of publication)? And is the availability of the software and any restrictions on use clearly stated in the manuscript?

The website is available in Beta right now. I does 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 reasaonable.

5. Does the manuscript clearly describe the problem the software is designed to address
Yes, they are trying to generate a rank ordered list of candidate genes for a disease. They use curated summary sites, and do not attempt to integrate other sources of data such as microarrays in GEO as another source of genes related to diseases. They don’t use the primary literature for text mining, but rely on curated databases which is fine, but limiting.

6. Does the manuscript clearly describe how the software is implemented?

Yes. It is pretty straightforward.

7. Does the manuscript clearly describe how the software performs and its advantages / limitations over existing applications?

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.

8. Does the manuscript state the software’s operating requirements

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.

9. Are the discussion and conclusions of the manuscript well balanced and adequately supported by the data?

Yes.

10. Do the title and abstract of the manuscript accurately convey what has been found?

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.

11. Is the writing acceptable?

There are a few too many “due to” for my tastes, but the writing is clear.

**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:

No competing interests.