Author's response to reviews

Title: Data mining of high density genomic variant data for prediction of Alzheimer's disease risk.

Authors:

Natalia Briones (natalia.briones@asu.edu)
Valentin Dinu (valentin.dinu@asu.edu)

Version: 3 Date: 2 December 2011

Author's response to reviews: see over
We thank the referees for their review of our manuscript and their constructive comments. Our responses to each reviewer’s comments are below; the reviewers’ comments are in *italics*, and our responses are in plain text.

MS: 3280416775772777
Research
Data mining of high density genomic variant data for prediction of Alzheimer's disease risk.
Natalia Briones and Valentin Dinu
BMC Medical Genetics

Referee number 1:

Major compulsory revisions:

1) *The manuscript is not written clearly which needs a major revision. For example, both Background and Methods sections should be rewritten extensively because I cannot find novelty in their proposed methods based on current version. Authors should provide sufficient references and description of state-of-the-art computational methods that are used now to perform the same task, which is obviously missing in the manuscript. The subsection “Analytical approaches for Alzheimer’s Disease association analysis” should be moved out of Background section because it does not fit there.*

Response:
The manuscript has been thoroughly revised to highlight more our findings regarding the genetics of Alzheimer’s disease than the methodology used. We moved the subsection “Analytical approaches for Alzheimer’s Disease association analysis” from the Background section to the Methods section. We trimmed and refocused the Background section and present the purpose of our study from the perspective of the recent genetic variants discovered and their putative epistasis done using GWAS.

2) *The presentation of manuscript makes me feel confused too, is this a methodology paper (suggests a new method) or research article (provide new predictions with experimental evidence)? The manuscript seems not belong to either of the above-mentioned two categories because authors used public available tools (PLINK and WEKA) to perform the data analysis such as logistic regression and random forest, and the predictions are not supported by other experimental evidences. These drawbacks make me feel difficult to access the quality of their study and predictions.*

Response:
We are presenting this study of a publicly available Alzheimer’s disease data set as a research article. We found new Single Nucleotide Polymorphisms associated with genes related to pathways that may be helpful for explaining the genetics of Alzheimer’s
disease and utilized them in multi-locus analysis for possible disease prediction and further disease understanding. We devised two different approaches from the ones originally utilized in the first publication of this data. Multi-locus disease association analysis is a complex, open-research problem, and we present two alternatives to approaching such analyses. The data is composed of one discovery group and two replication groups and our analysis was confirmed by 10-fold cross validation where there is testing on data that was not used for training; thus, the bias from the reported error rates was removed.

Referee number 2:

1. Authors put too much in Background section (8 pages). It'd better to reduce them to precisely reflect the subjects.

Response:
The manuscript has been thoroughly revised to highlight more our findings regarding the genetics of Alzheimer’s disease instead of the methodology used. We moved the subsection “Analytical approaches for Alzheimer’s Disease association analysis” from the Background section to the Methods section. We trimmed and refocused the Background section and present the purpose of our study from the perspective of the recent genetic variants discovered and their putative epistasis done using GWAS.

2. The comparison between two approaches is only at the description level, but showing one figure or table actually to better show the comparison.

Response:
Sub-sections for the two steps under the two approaches sections were added to clarify the sequence of the analysis process.

3. As the authors pointed out the approaches will improve the results from the analysis of another AD data set, It’d better to test it in this study in order to demonstrate the validity of the approaches.

This is a good suggestion, and we could request access to obtain additional Alzheimer’s data sets from NCBI’s database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap) for future testing and directions; however, for this paper we use cross validation which does testing on data that was not used for training, so that effectively removes the bias from the percentage values reported as error rates. The Cross-Validation WEKA algorithm randomly organizes the dataset and divides it into n folds of same size. In each repetition of the algorithm, one fold is used for testing and the other n-1 folds are used for training the classifier. An average over all folds is the result reported. The data is composed of one discovery group and two replication groups and our analysis was confirmed by 10-fold cross validation where testing is performed on
data that was not used for training; thus, the bias from the reported error rates was removed.