Reviewer's report

Title: A comparison of genomic profiles of complex diseases under different models

Version: 2 Date: 4 July 2015

Reviewer: Manish Datt

Reviewer's report:

Comments:
In the manuscript “A comparison of genomic profiles of complex diseases under different models” by Potenciano et al an attempt has been made to develop a predictive model to access individual predisposition to different complex diseases. The authors have performed a detailed analysis for seven diseases using five different machine learning algorithms with genotype and haplotype data as input features. For two out of seven diseases the predictive model based on boosting algorithm has high statistical accuracy. Overall, the study design is straightforward; however, the manuscript needs to be comprehensively revised for publication in BMC Medical Genomics.

Major essential revisions:
1) Analyses of 7 diseases show that for 2 of the disease (type 1 diabetes and rheumatoid arthritis) the boosting algorithm shows good performance. Curiously, both these disease are autoimmune diseases so authors should provide details about whether there was any overlap between the input features for these two diseases. Further, it would be interesting to see how this machine learning method performs in case of other autoimmune diseases (e.g. multiple sclerosis).
2) Based on the comparative analysis of different classifiers, the authors conclude that the boosting algorithm offers a “promising approach” for model building. However, the in case of bipolar disorder. The performance of complex machine learning methods (including boosting algorithm) is poorer than simple classifiers such as logistic regression and Bayes classifier. Therefore, it would be unjust to categorically present boosting algorithm as the best approach. It seems the predictive performance of the various statistical models would vary for different diseases. The authors should appropriately revise the manuscript to address this point.

Minor essential revisions:
1) The introduction section is too lengthy and needs to be shortened. In particular, there is repetition for description of different statistical measures (sensitivity, specificity, accuracy etc.) in the introduction and methods sections.
2) The quality of figure 1 is rather poor; the authors should make this figure in high resolution. Similarly, the resolution for Figure 5 needs improvement. Font size for axis labels and legends for figures 2-4 should be increased. Also, the
figures provided for review are black and white but there legend mentions about
the color lines denoting different AUC. Please make sure that color figures are
uploaded along with the manuscript.

3) The authors should describe Figure 5 in the manuscript. Simply writing “Figure
5 summarizes the whole procedure of our haplotype-based approach” at the end
of discussion is not sufficient.

4) The manuscript has 5 figures but the legends are provided only for 4 of these
figures.

5) The authors have compared the performance of different haplotype models for
different disease (Figure 2-4). A detailed comparison of different haplotype
models for a particular disease should also be done and discussed in the
manuscript. E.g. the authors should compare the performance of additive,
dominant, and recessive model for BD (and other 6 disease) and comment
whether a particular model (out of the 3) performs better than others or all the 3
have similar performance values.

6) “machine learning field within very different approaches able to build more
complex models from data” (Page 14) – What are the different approaches? It
would be nice to have a table having comparison between different approaches.

7) In the abstract – the full-form of the acronyms (such as WTCCC and AUC)
should be given.

8) In the second paragraph of introduction section there is redundant use of
dashes (-) these should be removed.

9) In methods section, “we had genome-wide SNPs genotyped for 1868,1988,
1748, 1952, 1860, 1963 and 1924 individuals respectively.” The sentence needs
to be re-written to make it clear that how many samples are selected for which
disease.

10) The author should copyedit the paper to improve the style of written English
to avoid syntactical errors and typos some these are listed below:
   “performed by [2]”: Should be written as performed by Evans et al [2].
   “cases a controls”: Cases and controls
   “range of them”: Range of p-values
   “correct pct”: What is pct?

11) The authors should avoid frequent use phrases such as “state-of-the-art” –
twice in one sentence (Page 14).

12) Table 7 and 8: In scientific notation use minus sign (-) with no spaces e.g.
1e-5 and not 1e – 5. Use scientific notation for all values in p-value column.

13) In various places in the manuscript, comma has been used instead of
decimal e.g. Page 3: 0,72 should be 0.72. Similar error needs to be corrected on
pages 7 and 9.

14) Under Haplotype-based predictors: First paragraph seems redundant; the
authors should rewrite this paragraph so that it aligns with the second paragraph.
15) In acknowledgement section, spellings of written should be corrected.
16) Supplementary material has a total of 89 tables and only Table S1 has been cited in the main manuscript. The main manuscript should have details about the content of supplementary material. The table in supplementary material needs to be reformatted so that they do not get truncated at the end of the page. Also, “P-value” should be written as “p-value” throughout the manuscript and supplementary material.
17) URL given at the end of end of reference 41 should be removed.

Discretionary revision:
1. “The median accuracy, precision, specificity, (Page 15)” – In addition to median, the range for values for these parameters should also be reported.
2. The authors should add a table having details of the 7 diseases (such as cause, symptoms, genes known to be involved, etc.)

**Level of interest:** An article whose findings are important to those with closely related research interests

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

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests