Reviewer’s report

Title: Genome-wide association studies reveal that the set of disease-associated alleles is enriched for derived low frequency alleles relative to HapMap and neutral expectations

Version: 2 Date: 16 July 2010

Reviewer: Nic Timpson

Reviewer's report:

Review of:

Genome-wide association studies reveal that disease-associated alleles are enriched for derived low frequency alleles relative to HapMap and neutral expectations.

Joseph Lachance

Comments to the authors:

The author has provided an amended version of the original paper and addressed all comments to reviewers. Most of these are ok (this is an interesting paper) and cover the alterations needed, however some require further attention and are outlined below along with comments on the new draft of the paper:

Response to rebuttal:

(i) 1000 genomes data IS now available and would be a valuable addition to this paper. The platforms covered are biased in the way of common variation and in efforts to attain a comprehensive assessment of evidence for selection in the regions of interest ("hits"), one would want to employ ALL known variants, or at least all of those known from public data. Imputation to 1000 genomes data density and performance of test for selection would be an important addition to this work and a genuine contribution to analyses of this kind.

(ii) Point (i) would aid the response to reviewer 1's comment in that there needs to be a clear outline of the contribution of this paper to the field.

(iii) With reference to the point on the qualification of "hits" and the new addition of a definition of this to the "empirical data" section - I have concerns as to the use (or at least a consideration of) the use of p\(^*\)10\(^{-5}\) as a threshold. This will allow for the inclusion of both false negatives and positives and will introduce a stochastic element to your appreciation of whether patterns in the selected regions are different or not in comparison to genomewide patterns under a null model. Hopefully, the effect of this will just be to introduce noise (assuming that there is an even distribution of false positives and negatives). However if there is a bias in either direction (just as there is for allele frequency in this work), this will influence the results and their interpretation. At the very least, one could perform
a sensitivity analysis where the threshold was higher (lower p values to that considered reliable at a genomewide level) - are the overall patterns the same here (and how does the incorporation of 1000 genomes data influence results at these thresholds??)?

(iv) Where you have performed a sensitivity analysis for the incorporation of allele frequencies from HAPMAP - please tell the readers that you have done this...

Response new version of the paper:

(i) Again - the title still suggests that the coverage of the analysis is of alleles of all frequencies - I think that (without the analysis of sequence data or otherwise) this has to recognise the coverage limitations of the work. Could achieve this just by stating that these are "genomewide association study derived disease associated alleles".

(ii) The point above is also reflected in the opening parts of the abstract in that you claim to be comparing results to the "rest of the genome". This is true, but only to the extent of a window of allele frequency and in certain populations. I think that this should be addressed.

(iii) Similar point for the conclusions of the abstract.

(iv) To some extent the tone of the intro could be changed. GWA studies have been VERY efficient at mining a portion of the variation in measured traits that is explicable to common variants of small effect. It is that these methods are only mining one part of the genomic architecture of disease that is restricting the explanation of "heritability" - not that they have "failed".

(v) To address the issue of whether the SNP set was random, is it possible to resample and compare by means test…?

(vi) Where a SNP was associated with several diseases, the author states that the mean allele frequency is taken across the studies. Does this take into account population of study (which will influence allele frequency and signal even within Europeans).

(vii) It is still not entirely clear how differences in allele frequencies are actually tested.

(viii) I would encourage a resume of main results at the start of the discussion - this can then be followed with specifically addressing points raised by the study.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.