Reviewer’s report

Title: Genome-wide association studies reveal that the set of disease-associated alleles is enriched for derived low frequency alleles

Version: 1 Date: 28 April 2010

Reviewer: Nicholas Timpson

Reviewer’s report:

The author provides a well written MS which has the impression of sure understanding in the field of population genetics. An interesting topic, in the characteristics of genomewide association study derived effect alleles, is explored and this is done in a comprehensive manner (though is missing some important references work which has already been undertaken in this field). Key elements in a paper and analysis such as this are the derivation of comparator groups and the undertaking of appropriate analyses. These are (in the most part) well explained and seem to deliver predictable, though interesting, results. However, there are considerable alterations that will be needed before this paper can be considered ready for publication and there are also a series of key questions which need to be addressed. A detailed response to the paper follows:

(i) In the title of the paper, the term “low” is used to describe the characteristics of disease-associated alleles. This is a loaded term and should be avoided (at least before definition) if possible. The types of variants that are being assessed here are those derived from genomewide association study and by default (and owing to the capabilities of this approach to genetic analysis) will not be rare or have “low” frequencies. These may be relatively so, but in its current state, this may be misleading.

(ii) Again, in the results section of the abstract, a loaded term is used. Here an excess of “rare derived alleles” is referred to and whilst correct in a relative sense here, is not an expression that can be used without definition. If possible, this should be reworded.

(iii) Throughout the document, please refrain from using the term “significant” to describe notable statistical patterns. Rather, please just comment on the presence or absence of difference or effect and then qualify it in terms of the accuracy and precision of the data.

(iv) It is not clear in this paper what qualifies a genomewide association study “hit”. The database used is indeed a reliable collection of results from this type of study. However, the authors should be transparent about the breadth of association study signals that have been included in these analyses from an early stage.

(v) Previous studies relating genetic association signals to the rest of the genome (in terms of global characteristics) and reporting the nature of genomewide
association studies both within and across populations have been reported. They have not been on a scale such as this (i.e. not using the repository of genomewide association results used here), but should be included as a context for this study.

(vi) The first line of the “Background” can be removed and indeed in the second line, there should probably be reference to the WTCCC landmark paper of 2007.

(vii) The nature of available genomewide association study findings (i.e. common for the main and small in effect) and the lack of explanation of all phenotypic variance does not preclude the validity of the common disease/common variant hypothesis. By their very nature, we do not expect common variants to have large effects (owing to their fitness burden), hence the validity or not of this hypothesis (or others) remains entirely subject to the nature of findings not yet made. The remaining variance may be explained by further common variants with even smaller effect – as is being show in some ongoing, large scale research) or more rare variants with larger effects. This remains to be seen and as such the statements on this in the background should reflect this.

(viii) It is somewhat difficult to see the need for the second paragraph of the background.

(ix) With this, the third paragraph classifies alleles (frequency/ancestral-derived/effects), however this is again not really contributing to the paper. These classifications do not really appear later on in the paper and other than the definition of “common” and “rare” in this type of study (i.e. not syndromal research), these details may not be needed. If this section is to be kept, it would need shortening considerably and incorporating into the drive of the paper more directly.

(x) The last paragraph of the background (which I assume should follow as per a normal introduction) contains aspect which can be moved both into the methods and results sections. There is no real summary of the aims of this work which is what I would expect for this section.

(xi) The methods section was clear and I understood the ordering of the explanations, however this took two reads and I think for the more general readership of BMC Med Genome will need to be simplified where possible. At the outset of each section, it would be good to include a brief sentence as to the reason for each new part of the analysis. If this follows the same structure of the aims section at the back of the introductory stages and is picked up in the discussion, this should make the paper simpler at first read. For example, in the first instance I questioned the use of unweighted null distribution calculation. However you then go on to describe how this is augmented for use. A summary of the overall aim and then its description or a running statement across the methods descriptions may help to avoid this.

(xii) Were the 1000 SNPs selected for HAPMAP for the derivation of null distributions independent? LD is stated to be “minimal”, however the extent of the
LD between these variants is critical to the derivation of this distribution.

(xiii) Were outgroup allelic states available for all variants including the genomewide association results?

(xiv) Please refrain from using the device “Note that…” in prose. Either use a proper footnote or include these sections into the main text.

(xv) As mentioned above, clarification of the sections included in the methods is important, however there is a fine line between the explanation of population genetic theory and the description of the methods used. Whilst it may be difficult, it is important to remove as much of the theory as possible and to simply present a possible reference to the aim of each section and the methods used in that part of the work.

(xvi) Please replace the phrase “one over” with either a formulaic presentation of the expression or write it in full.

(xvii) Characteristics of the “risk” alleles are taken from the studies that derived them. However, given the possible biases and heterogeneity that the author recognizes in the collection of data in this way, why not just flag the SNPs and alleles with genomewide study database information and then use HAPMAP data for their assessment. This would seem to avoid the problems mentioned in the “empirical data” section.

(xviii) Comments on the effective population size of humans seem to have awkward wording. This section also needs a clear indication of its contribution to the main drive of the paper.

(xix) In the results section, the comments made earlier with respect to the use of terms like “low” and to frequencies such as “0.39” need to be contextualized. (This is reflected in the author’s own comments in the conclusion “low hanging fruit”!)

(xx) Many of the results (i.e. disease variants being more rare, younger and derived) are as one would expect from population genetic models. This should be referred to (in light of the confirmatory findings here).

(xxii) In the second paragraph of the “natural selection…” part of the discussion, the sentence “that diseases with late onset are less likely to be found at low frequencies.” Could be made a lot more clear by just stating that “that diseases with late onset are likely to be found at higher frequencies.”

(xxii) In the discussion the question “are causal disease alleles more likely to be ancestral…” – this has to be framed in light of the trait of interest… the direction of the answer to this will be completely dependent on which trait is analysed.