Author's response to reviews

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

Authors:

Joseph Lachance (Joseph.Lachance@sunysb.edu)

Version: 2 Date: 25 June 2010

Author's response to reviews: see over
First, I would like to thank both reviewers for the quality and quantity of their comments. I have made an earnest effort to address each of their suggestions and believe that the changes made have substantially improved this paper (see responses below). A number of additional analyses were done: Characteristics of GWAS alleles were compared for different dates of publication, genotyping platforms, and number of SNPs genotyped. In addition, I tested whether exclusion of GWAS data from studies of non-European individuals had a major effect. Characteristics of loci (as opposed to alleles) associated with genetic disease were also analyzed. This analysis involved comparisons between derived frequency distributions of HapMap SNPs and disease-associated SNPs. The numbering of Figure 4 and Figure 5 was switched in this revision, and derived frequency distributions were included as a panel in the renumbered Figure 4. The order of the “Estimated ages of SNPs” and “Odds ratio” subsections was also switched. I hope that the improvements made are sufficient enough for this paper to warrant publication in BMC Medical Genomics.

Sincerely,
Joseph Lachance

Graduate Program in Genetics
Department of Ecology and Evolution
Stony Brook University

RESPONSE TO REVIEWERS

Reviewer: Jing Hua Zhao

Major Compulsory Revisions
I feel the work is likely to be consolidated in several ways. As the comparison was designed to be in line with SNPs reported from GWASs, a random set of 1,000 SNPs was selected from HapMap; could there be selection bias?

I do not believe that there is a selection bias in the 1000 HapMap SNPs (mainly due to the fact that these SNPs were selected via a random number generator and they were found to have minimal LD). That said, I downloaded an additional 1000 random SNPs from the HapMap (CEU) and Perlegen (EUR). Mean frequencies of minor alleles were roughly similar for each of these datasets (HapMap SNPs in this paper: 0.219, additional HapMap SNPs: 0.214, Perlegen SNPs: 0.199). Also, percentages of derived HapMap alleles in this paper were comparable to whole genome estimates from the recent Neanderthal genome paper (Green et al. 2010). In any case, additional details of how SNPs were selected are now mentioned in the “Null expectations: HapMap data”
subsection. As an additional control, disease-associated alleles were analyzed for different genotyping platforms (see Table 2). Neither manufacturer nor number of genotyped SNPs had a major effect on the characteristics of disease-associated alleles, suggesting that there was not a major ascertainment bias.

It is known that most GWASs, as with those listed in the NHGRI, has also been selected largely for common variants … It is preferable with the availability of HapMap III/1000genomes data, a better picture can be drawn.

I too am looking forward to the 1000 Genomes data. It will be really interesting to see is how much characteristics of SNPs vary from population to population (particularly in Africa). This is likely to have implications for future studies (GWAS or otherwise), as methods that work well for the alleles segregating in one population may not be ideal for another population. However, population-level comparisons are beyond the scope of this study. The HapMap build used in this study contained merged data from phase II and Phase III, and this details is now included in the “Null expectations: HapMap data” subsection.

The author made no differentiation between different GWASs which themselves may contain false positives. it might be helpful to extend the consideration of allelic characteristics to loci or genes as these are often characterised in GWASs.

The set of all genome wide association studies is quite heterogeneous and I agree with the reviewer that many of these may contain false positives. In addition, studies vary in terms of ascertainment bias, sample size, and admixture in case/control studies. Controlling for all of these issues in a meta-analysis is impossible (especially since the degree to which each of these issues affects a particular study is unknown). Thankfully, the sheer number of studies analyzed in this paper is likely to reduce the effect of any one study. There is also no a priori reason to expect false positives to cause the set of all disease-associated alleles to contain more low frequency derived alleles than either null expectation. In addition, the fact that the characteristics of replicated alleles resemble the overall set of disease-associated alleles (Figure 3 and Table 1), suggests that the patterns observed in this study are real.

The above comment by the reviewer prompted me to analyze the characteristics of loci (rather than alleles) that show statistical associations with genetic disease. This is possible via comparisons of derived frequency distributions of HapMap SNPs and disease-associated SNPs. The details of this analysis are included in the “Disease-associated alleles” subsection. Interestingly, the loci involved tended to be similar (mean frequency of derived alleles was 0.426 for disease-associated SNPs and 0.432 for HapMap SNPs). This suggests that much of the difference between disease-associated alleles and the rest of the genome was due to properties of alleles rather than loci. Derived frequency distributions of HapMap and GWAS SNPs are now featured in Figure 4.
As an additional control, I excluded disease-associated alleles from studies that did not involve European populations and re-analyzed the data. Exclusion of this data had only a minor effect (see “Disease-associated alleles” subsection). In addition, Hindorff et al. (2009) mention that GWAS alleles are overrepresented in nonsynonymous sites and promoter regions, and this paper is now cited in the introduction.

A further but perhaps minor limitation is the author’s decision to refrain from consideration of multiple mutations.

The vast majority of SNPs are biallelic and per nucleotide mutation rates are quite low, which suggests that multiple mutations should not be a major issue. In any case, I have changed the wording in the discussion to reflect the fact that multiple causal mutations can appear in multiple genetic backgrounds, but this need not be the case.

Finally, it would strengthen the paper by highlighting methodological developments made in the paper.

Additional sentences explaining why particular methods were used are now included in the manuscript (as per the suggestions of both reviewers). It might not directly relate to the question of this reviewer, but other studies have looked at derived vs. ancestral allele patterns for specific diseases. For example, 6 of 17 alleles associated with an increased risk of diabetes were derived alleles (Southam et al 2009, Diabetologia).

Minor Essential comments
Abstract
Can some further elicitation be given in Background, regarding the utility of addressing the “open question is ... ” (some can be seen in Background)? Aspects related to effect size in Results are known in the literature. In Conclusion it might worthwhile to indicate that the investigation is more specific to common variants.

An additional sentence describing the utility of addressing the open question is also included. The fact that low odds ratios confirm previous findings is now mentioned in the abstract.

Background
As mentioned earlier, it is helpful to note that a variety of GWAS consortia is actively working on rare variants. It is interesting to compare results to earlier work [8] given more SNPs were used in current investigation.

Agreed, the analysis in this paper now examines whether the characteristics of disease-associated alleles differ for early vs. later studies. Somewhat surprisingly, there was not
that much of a difference. This may be due to the low power of GWAS to detect associations between disease and alleles at frequencies < 0.05 or > 0.95. While later studies presumably include more SNPs with rare alleles, the power to detect associations with these alleles is likely to be low.

Methods
It might (not??) always be true with “important difference between disease-associated alleles and SNPs from HapMap project is that the latter are not weighted by their probability of detection”. For instance, we have used a case-cohort design in which the subcohort sample is representative of the population under consideration.

This sentence has been reworded for clarity.

The earlier assumption was that “allele characterized in this paper are marker alleles”, but in the actually investigation a complete linkage [disequilibrium] between causal and marker loci was assumed. Is this consistent?

Yes. The important consideration here is that all copies of the causal allele appear on the same (marker) genetic background. Allele frequencies may differ at causal and marker loci. In any case, the mention of complete linkage has been removed from the “Null expectations: neutral theory” subsection. So long as neutrality is assumed, the allele frequency distribution of marker alleles should be independent of linkage phase and allele frequency at the causal locus.

Please revise “..., complete linkage, multiplicative dominance”. The statistical power as a function of MAFs, ORs, etc., esp. when in complete LD, is well recognised in the literature, perhaps will be more appropriate as supplementary material/appendix.

This material has now been moved to the appendix, and the equations have been renumbered.

Could the author provide the rationale over the modelling, e.g., the mathematical simplicity in (7) being the expected/coupling probability?

The equation mentioned by the reviewer has been renumbered Equation 6. It is purely coincidental that the expression in Equation 6 is equal to the expected heterozygosity. Different sample sizes and/or odds ratios would yield a different expression. Details regarding this point have been added to the main text. Interestingly, “x*(1-x)” does appear in equations quantifying linkage disequilibrium (see Appendix).
Results
Could the author give more details about the chi-squared test with 19df? Could it be possible that tests such as Kolmogorov-Smirnov perform better?

Frequency distributions include ancestral vs. derived states of alleles. Because of this, the data are not continuous and the Kolmogorov-Smirnov test seems inappropriate. The expected number of alleles in each bin was found by multiplying the total number of observed alleles by the expected proportion of alleles occupying that bin. For example, the weighted neutral theory expectation is that 2.05% of alleles will be ancestral and have a frequency between 0.10 and 0.20. A total of 1143 GWAS alleles were observed. The expected number of alleles for this bin = 23.4 (0.0205 * 1143), compared to the 51 GWAS alleles that were actually observed for this bin. There are 20 total bins (10 bins for ancestral alleles and 10 bins for derived alleles). This gives 20-1=19 degrees of freedom.

Discussion
As commented earlier, the lists at NHGRI may have been selected in terms of MAFs. It is understood that apart from larger sample size, finer map is also necessary for future success in GWASs. The relative inability of GWAS to explain the high inheritability has been attributed to a variety of sources such as rare SNPs, structural and epigenetic variants, or multiple alleles with additive effects or synergistic genetic interactions (Dowell et al 2010, Science 328:469), as with gene-environment interactions.

An additional sentence mentioning other possible causes of the “missing heritability” has been added to the first subsection of the Discussion. The benefits of increasing the number of genotyped SNPs were also mentioned. As an aside, I’m definitely in the “epistatic interactions are important” camp.

Besides this investigation, consideration of genomewide data other than those from GWAS or HapMap might facilitate a good picture. In terms of allelic age, it is known that generally commoner polymorphisms are old. These might have been benign or beneficial.

It would definitely be interesting to see if the characteristics of alleles in the HGDP differ from HapMap/neutral/GWAS alleles. Do these patterns differ from population to population? The effects of human migration and serial bottlenecks might produce interesting patterns, especially with respect to the presence of high frequency derived alleles. However, this is beyond the scope of the current study. In any case, the link between SNP age and fitness is now mentioned in the “Natural selection against deleterious alleles” subsection.
Reviewer: Nicholas Timpson Reviewer's report:
(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.

With the above suggestion in mind, I have revised the title to read: “Genome-wide association studies reveal that disease-associated alleles are enriched for derived low frequency alleles relative to HapMap and neutral expectations”

(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.

The wording has been changed (making things less ambiguous/loaded).

(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.

Following the advise of this reviewer, I have removed many uses of the term “significant” from this paper.

(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.
A sentence that describes the criteria for inclusion of an association is now included in the “Empirical data” subsection of the Methods.

(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.

An additional sentence describing previous findings has been added to the background section.

(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.

The first line has been removed and the suggested reference is now included.

(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.

Agreed. Details relating to this point have been added to the “Odds ratios” and “Statistical power, sample sizes, and allele frequencies” subsections.

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

I have removed this paragraph. However, two sentences that made the distinction between causal and marker alleles were retained and moved to the next paragraph.

(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.

I believe that this paragraph needs to remain, as it lays the groundwork for the results section. In any case, this paragraph has been greatly shortened (mention of minor allele frequencies has been removed, among other changes).

(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.

A sentence that describes the aim of this work has been added to the last paragraph of the background. Also, this paragraph has been shortened.

(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.

Thanks. This is a great suggestion and each of the Methods subsections now include a sentence explaining why a particular type of analysis was used.

(xii) Were the 1000 SNPs selected from 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.

Yes, these SNPs were independent. Each SNP was at least 200kb away from other randomly selected SNPs.

(xiii) Were outgroup allelic states available for all variants including the genomewide association results?
Yes. In the event that an outgroup allele was unavailable that allele was omitted from the dataset. This point is now mentioned in the methods section.

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

Sentences beginning with “Note that” have been rewritten.

(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.

I was able to trim a few sentences from the theoretical parts of the Methods section. Equations of statistical power are now found in an appendix, and descriptions of effective population size have been removed.

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

This sentence has been reworded.

(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.

An interesting idea… Out of curiosity I did as the reviewer suggested, and found that allele frequencies were quite similar in both cases. Under the original methodology: GWAS alleles had a mean frequency of 0.394 and a median frequency of 0.360. Using only CEU HapMap frequencies: the mean frequency was 0.405 and median frequency was 0.368. Because of the lack of major differences, I chose to retain the old methodology. As an additional control, I excluded disease-associated alleles from studies that did not involve European populations and re-analyzed the data. Exclusion of this data had only a minor effect (see “Disease-associated alleles” subsection).

(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.

The wording in this section has been cleaned up. Also, the specific effective population size estimate for humanity has been removed.

(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”!)

An effort has been made to reduce the ambiguity of terms like “low”. Where possible, differences between GWAS alleles and null expectations were quantified. I kept the “low hanging fruit” wording in the Conclusion. Hopefully easily detectable associations will not be confused with associations involving low frequency alleles!

(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).

Matches between results and population genetic theory are now mentioned.

(xxi) 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.”

Thanks. The wording of this sentence has been improved as suggested.

(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.

Agreed. Perhaps the causal alleles for many neurological diseases will turn out to be ancestral (indicating that what we view as “disease” is the character state shared with our recent common ancestors). In any case, I have slightly changed the wording of this sentence to indicate that the answer to this question is likely to be trait-specific.

ADDITIONAL DETAILS ADDED:
Percentages of derived HapMap alleles were comparable to whole genome estimates from the recent Neanderthal genome paper (Green et al. 2010). This detail was added to the “null expectations” subsection.

Added a sentence mentioning possible ascertainment biases of HapMap SNPs to the “Estimated ages of SNPs” subsection.

Future studies may benefit from the inclusion of many SNPs with low frequency derived alleles. This detail is now mentioned in the “Statistical power, sample sizes, and allele frequencies” subsection.

A sentence mentioning population heterogeneity was added to the “Genetic background and linkage phase…” subsection.

Added a sentence mentioning genetic linkage analysis to the conclusion.

Changed the wording in Table 1 from “Probability(ancestral)” to “Proportion ancestral”.

Four (out of 1147) GWAS alleles were found to be duplicated in the dataset, and Table 1 and details in the main text have been (slightly changed) as a result.