Author's response to reviews

Title: Discovery and Replication of SNP-SNP Interactions for Quantitative Lipid Traits in over 60,000 Individuals

Authors:

Emily Holzinger (emily.holzinger@nih.gov)
Shefali Verma (szs14@psu.edu)
Carrie Moore (carrie.c.moore@dm.duke.edu)
Molly Hall (hallma@mail.med.upenn.edu)
Rishika De (rishde@gmail.com)
Diane Gilbert-Diamond (diane.gilbert-diamond@dartmouth.edu)
Matthew Lanktree (mlanktree2012@meds.uwo.ca)
Nathan Pankratz (pankr018@umn.edu)
Antoinette Amuzu (Antoinette.amuzu@lshtm.ac.uk)
Amber Burt (aaburt@uw.edu)
Caroline Dale (Caroline.dale@lshtm.ac.uk)
Scott Dudek (sud23@psu.edu)
Clement Furlong (clem@uw.edu)
Tom Gaunt (Tom.gaunt@bristol.ac.uk)
Daniel Kim (bstk@uw.edu)
Helene Riess (thorand@hemholtz-muenchen.de)
Suthesh Sivapalaratnam (s.sivapalaratnam@amc.uva.nl)
Vinicius Tragante (v.tragante@umcutrecht.nl)
Erik van Iperen (e.p.vaniperen@amc.uva.nl)
Ariel Brautbar (brautbar@bcm.edu)
David Carrell (carrell.d@ghc.org)
David Crosslin (avid.crosslin@gmail.com)
Gail Jarvik (pair@uw.edu)
Helena Kuivaniemi (siskohkuivan@gmail.com)
Iftikhar Kullo (kullo.iftikhar@mayo.edu)
Eric Larson (Larson.e@ghc.org)
Laura Rasmussen-Torvik (ljrtorvik@northwestern.edu)
Gerard Tromp (trompgc@gmail.com)
Jens Baumert (baumert@helmholtz-muenchen.de)
Karen Cruickshanks (cruickshanks@episense.wisc.edu)
Martin Farrall (mfarrall@well.ox.ac.uk)
Aroon Hingorani (a.hingorani@ucl.ac.uk)
G.K. Hovingh (g.k.hovingh@amc.uva.nl)
Marcus Kleber (marcus.kleber@medma.uni-heidelberg.de)
Barbara Klein (kleinb@epi.ophth.wisc.edu)
Ronald Klein (kleinr@epi.wisc.edu)
Wolfgang Koenig (wolfgang.koenig@uniklinik-ulm.de)
Kari North (kari_north@unc.ed)
N. Charlotte Onland-Moret (n.c.onland@umcutrecht.nl)
Alex Reiner (apreiner@u.washington.edu)
Philippa Talmud (p.talmud@ucl.ac.uk)
Yvonne van der Schouw (y.t.vanderschouw@umcutrecht.nl)
James Wilson (jgwilson2@umc.edu)
Reviewer reports:

Reviewer #1: The manuscript by Holzinger et al. describes a discovery/replication approach for identifying SNP-SNP interactions related to lipid variation in a large cohort of 60,000 individuals. It is thought that SNP-SNP interactions play an important biological role, but finding them is difficult due to the large numbers of SNP combinations that require testing. The purpose of the manuscript was two-fold: 1) reveal novel insights into the genetic etiology of lipid levels, and 2) develop a pipeline for computationally efficient interaction analysis. The paper did a very good job of addressing the 2nd goal; however, I think additional attention should be paid to describing the novel insights into the genetic etiology, and how these findings fit into the existing literature. Overall, this manuscript is well written and the methods will be impactful to the field.

First, we would like to thank the reviewer for taking the time to provide insightful and helpful feedback on our manuscript. We have carefully reviewed each of the comments, and have made significant improvements to our paper accordingly. Below, we detail how each of the comments were addressed and the specific changes that were made in the manuscript.

Major Revisions:

1. Page 11, line 268-269 of the methods, the authors state '…individuals of self-reported European ancestry, subsequently verified using principal component analyses.' How was this verified? Were the clustered with other mixed race cohorts or was this verified by only clustering this group?

For our study, individuals with self-reported European ancestry were verified by showing that they clustered with the corresponding HapMap panel using principal components analysis. Those that self-reported European-American and that clustered with the CEU HapMap individuals were
selected for this study. This was done for all of the discovery and replication cohorts. We have clarified this by editing the following sentence in the methods section of the manuscript:

All of the individuals in our analysis were self-reported European ancestry, subsequently verified using principal component analyses by selecting individuals that clustered with the CEU panel from HapMap, and > 21 years of age.

2. Page 12, line 296, "SNPs with a main effect p<0.001 based on a previous GWAS regression…” This seems like quite a liberal threshold, how was this selected?

We agree that the threshold of p<0.001 is very liberal for strictly assessing main effect significance. We had two specific motivations for selecting this threshold for our study: 1. to allow for interactions that may be present in the absence of large, genome-wide significant main effects, and 2. to reduce the SNP set to a size that allowed for a manageable exhaustive SNP-SNP interaction analysis. We have clarified our reasons for selecting this threshold by adding this information to the methods section of the manuscript (page 12, line 297).

3. Page 13, lines 309-314, the authors describe covariates that were included in the model. It is not clear whether these were selected to be included in the model (e.g. backwards selection) or whether they were forced into model. If they were forced into the model, were they all significantly associated with the lipid phenotypes?

The top SNPs selected for Main Effect filter interaction analysis were from a GWAS meta-analysis that adjusted for the same covariates. These covariates are known to have effects on lipid levels based on previous studies and what is known about the biology of lipid traits. In order to control for these effects and to remain consistent with the study from which these SNPs were chosen, we adjusted for the same covariates in our study. We have clarified our reasons for adjusting for these covariates by editing the following part of the methods section (page 13, line 312):

We adjusted for age, sex, smoking status, type 2 diabetes status, BMI, medication use (use or no use of lipid lowering drugs), and potential population substructure (top 10 principal components) by including these as covariate terms in the linear regression models for each of the four lipid traits. We included these covariates to control for any factors outside of genetics that may have an effect on lipid levels and to remain consistent with the previous GWAS from which the SNPs for the Main Effect filter analysis were chosen.

4. In the same section, it states that the top 10 PCs were included as covariates, was including 10 PCs necessary since these were all from self-reported European Ancestry (which was stated previously as being validated by PCA). How much variation was explained after the first PCs and are you wasting degrees of freedom by incorporating so many PCs?

We thank the reviewer for identifying a part of our methods section that needed further clarification. While the top 3 PCs did control for a substantial portion of the variation, we used the top 10 PCs to control for any residual variation, especially since this analysis was done across many cohorts in many different regions of the country. Further, as with the other
covariates, we wanted to remain consistent with the previous GWAS from which the main effect SNPs were chosen that also corrected for the top 10 PCs. We have clarified this by adding the following sentence to the manuscript (page 13, line 321):

We chose to include the top 10 principal components to remain consistent with the previous GWAS and to control for any residual variation as we were performing these analyses in a combined cohort that included individuals from various parts of the country.

5. Page 13, lines 317-318, "We adjusted the threshold based on the number of estimated independent models that were tested." How was this determined?

We adjusted the p-values for the discovery models selected for the replication analysis using the total number of SNP-SNP models that were tested. These were estimated to be independent based on the LD pruning for the main effect filter analysis and the SNPs being in different genes for the Biofilter analysis. We have re-worded this sentence to make this concept clearer (page 15, line 326):

We adjusted the threshold using a Bonferroni correction based on the total number of number of models that were tested for each filtering methods. We estimated these models to be independent due to the LD-pruning in the main effect filter analysis and the SNPs being in different genes for the Biofilter analysis (Fig 1).

6. P 14, lines 326-331, The purpose of the 'proxy models' are not clear. Can you provide the rationale for why these were generated?

We have added the following clarification for why we generated the proxy models for the replication phase of the analysis (page 14, line 341):

The purpose of these models was to capture signals in the replication data that may have been missed due to allele frequency differences between the discovery and replication cohorts.

7. Were all the replication cohorts also of European ancestry?

All of the cohorts used in this analysis were of European ancestry. We have clarified this by adding this information to the description of the replication cohorts in the methods section of the manuscript (page 15, line 362).

8. Page 14, lines 346-347, "Briefly, each cohort tested for population stratification and relatedness, adjusting accordingly." First of all, this sentence should be reworded,? However, more importantly, how was this performed? Was this done the same as for the discovery set?

We thank the reviewer for noting this wording error and the need for more detail here. For the replication cohorts, we only had access to the summary level data. All QC was done by analysts at the individual institutions that provided the data. Each institution performed assessments and corrections (when necessary) for both population stratification and relatedness. The exact methods that were used to perform these corrections varied across cohorts. We verified that that
methods used were appropriate and adequate for our study for each replication cohort. The full
details for each of these corrections can be found in the methods descriptions included in the
references provided in Table 3. We have corrected the wording error and we have clarified the
QC procedures by adding the following sentence to the manuscript (page 15, line 358):

In each of the replication cohorts, population stratification and relatedness were assessed and
adjusted for accordingly. The full details for the QC procedures can be found in the references
provided for each replication cohort in Table 3.

9. Page 16, line 387, Was the LD pruning only done for the main effect?

LD pruning was only done for the main effect filter analysis as these SNP-SNP models are
generated exhaustively will include SNPs that are in the same genes with potentially very high
LD. For the Biofilter analysis, the SNP-SNP models are specifically chosen using SNPs in two
different genes. We have clarified this by adding the following sentence to the manuscript (page
12, line 295):

No LD pruning was done for the Biofilter interaction analysis, as these models are specifically
generated using SNPs that are in different genes.

10. What are the biological/translational implications of these findings? The authors mention
the top replicating SNPs, but do not address the biological plausibility of these findings. For
instance, CETP has been previously documented as being associated with lipids, but this is not
mentioned in the discussion. Were any of the replicated interactions in genes that were
previously identified in single SNP association analyses? Are any of these interactions pointing
to new biology that would not be detected in a single SNP GWA?

We thank the author for identifying this lack of insight into the biological implications of our
findings. Many of our top findings were in genes known to impact lipid biology. Further, many
of the SNPs in the interaction models did have large main effects and/or have been identified
previously in other GWAS. One reason for this is that the SNPs from the IBC chip were selected
to interrogate regions of the genome more likely play a role in the genetics of cardiovascular
disease. For our study, we are specifically looking for interaction between the SNPs in these
genes outside of the main effects. To address this comment, we looked for previous studies that
have identified interactions within genes with known main effects. Interestingly, we did identify
one study that found an interaction between two of the SNPs in our top replicating models for
HDL (rs4783961 and rs1800775). These SNPs represent two of the independent sets of SNPs
shown in Figure 2. Both SNPs are in the promoter region of CETP. We added the reference to
the previous study and the following text to discuss why an interaction between these two
variants may correspond to changes in HDL-C levels (page 18, line 435):

Genetic interactions are often described as gene-gene interactions, and are usually studied by
specifically looking for variants in different genes that could be indicating novel pathways (e.g.
protein-protein interactions that have not been previously identified using genetic data).
However, intergenic interactions, such as those that we observed in this study, should not be
ignored, as they may contribute to a substantial proportion of the genetic architecture. Our top
replicating models for HDL-C consisted of two SNPs in CETP. Many of these model replicated across cohorts with the top replication p-value for the likelihood ratio test being 3.0 x 10-12 (Table 1 and S1 Table). LD patterns suggest that there are three independent sets of SNPs that represent many of the top models for the CETP-HDL associations. Further, many of these SNPs are in the promoter region of CETP. Most notably, a previous study identified a functional interaction between two of the SNPs in one of our top models (model 9: rs4783961 and rs1800775) that resulted in changes in CETP promoter activity [26]. As discussed in this study, this could be explained by shared transcription factors that may result in non-linear changes in CETP and HDL-C levels when the variants occur together. These results provide further support for studying intergenic non-linear effects and that they could be important for both accurate phenotype prediction and for understanding the function behind why specific variants in this gene have certain effects on HDL-C levels.

Minor Revisions:

1. Page 11, line 275, references Table 3. I believe this is correct to reference table 3, but this is the first table mentioned in the manuscript, so this should be reordered.

We thank the reviewer for identifying these errors in our manuscript. We have re-ordered the tables so that Table 1 is the first one that is mentioned in the paper.

2. Page 12, lines 293-294, "missing genotype rate > 95%". Should this be SNPs were removed if had >5% missingness? It seems unlikely that SNPs missing 94% would have been included, but this should be clarified.

We did remove SNPs with missing rate > 5%. We have corrected this mistake in the manuscript.

3. Page 14, line 346, I believe Table 1 is actually referring to Table 3. However, as described above, this should probably be reordered to be Table 1.

We have corrected both the order to the table and the reference to the incorrect table.