Author’s response to reviews
Title: GWAS and enrichment analyses of Non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE network

Authors:
Bahram Namjou (bahram.namjou@cchmc.org)
Todd Lingren (todd.lingren@cchmc.org)
Yongbo Huang (yongbo.huang@cchmc.org)
Sreeja Parameswaran (sreeja.parameswaran@cchmc.org)
Beth Cobb (beth.cobb@cchmc.org)
Ian Stanaway (bard@uw.edu)
John Connolly (ConnollyJ1@email.chop.edu)
Frank Mentch (MENTCHF@email.chop.edu)
Barbara Benoit (BBENOIT1@PARTNERS.ORG)
Xinnan Niu (xinnan.niu@vumc.org)
Wei-Qi Wei (wei-qii.wei@vumc.org)
Robert Carroll (robert.carroll@vumc.org)
Jennifer Pacheco (japacheco@northwestern.edu)
Isaac Harley (ISAAC.HARLEY@UCDENVER.EDU)
Senad Divanovic (Senad.Divanovic@cchmc.org)
David Carrell (david.s.carrell@kp.org)
Eric Larson (eric.b.larson@kp.org)
David Carey (djcarey@geisinger.edu)
Shefali Verma (szs14@psu.edu)
Marylyn Ritchie (marylyn@pennmedicine.upenn.edu)
Ali Gharavi (ag2239@cumc.columbia.edu)
Version: 1 Date: 30 Apr 2019

Author’s response to reviews:

April 30, 2019

Lin Lee
Editor-in-Chief
BMC Medicine

Dear Dr. Lee,

Thank you for the opportunity to submit the revised manuscript, BMED-D-19-00262: entitled “GWAS and enrichment analyses of Non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE network” for publication in BMC medicine under the special collection: “Beyond Big Data to new Biomedical and Health Data Science: moving to next century precision health”; Guest editor Dr. Marc S. Williams.

Our responses to the comments from the reviewer are presented below. All changes in the manuscript are highlighted in red text.
Response to Editor:

Both reviewers have mentioned that your manuscript would benefit from analysis confirming associations in an independent cohort. We would strongly suggest that you consider doing this. Further consideration of your manuscript will be dependent on the changes made.

Dear Editor: Thank you for your consideration and comment. Please note that in this study we evaluated more than 80,000 participants that have been genotyped independently in 10 medical centers across eMERGE network and all reported results were replicated consistently across network with no evidence of heterogeneity. Please also note that some of our results have been further replicated independently when we applied PheWAS approach conditioning on NAFLD status. While we agree with the reviewers that additional replication is desirable, the important issue is that we have a definitive phenotype (i.e., NAFLD activity scores from liver pathology) and that available replication cohorts such as UK Biobank do not have the necessary information (i.e., path reports) that would allow meaningful replication. Indeed using novel bioinformatics and natural language processing, it took us more than a year to fully extract all the necessary phenotype findings from electronic medical records of participants and characterize the phenotype. Below we addressed all of the critics raised by the reviewers:

Reviewer reports:

Reviewer #1 (Antonio Julià): The current study describes a genetic analysis on non-alcoholic fatty liver disease (NAFLD). Risk and Severity are analysed as study endpoints. Using Natural Language Processing, NAFLD cases are ascertained from the eMERGE network database. The associated GWAS data is imputed and used to test for association with risk and with susceptibility phenotypes associated with NAFLD severity. No new risk locus is found, but a few variants are associated at the genomewide scale with NAFLD-associated severity phenotypes. The latter genetic findings are not tested for replication in an independent cohort. Biological information is integrated into analyses to provide a more functional interpretation of the genetic basis of NAFLD.

Major comments

1) In Methods, the study cohort is defined as "9,677 European ancestry participants", but then PCA analysis shows evidence for African, European & Asian ancestries. This should be clarified. Then, the three main PCs are used to correct for population stratification. Given that
there are divergent ancestries in this study, it is expected that the main PCs will hold most of the variance. However, there could still be cryptic ancestry influencing the observed associations. A sensitivity analysis would help increase confidence on the observed association. The authors should include additional PC's (e.g. top 10) as covariates and see if the new reported associations hold. This is particularly relevant since there are no replication cohorts for the new findings.

Response: Thank you. As highlighted in the text, in this study, we only evaluate those with self-reported European ancestry and performed a secondary PC analyses only on those participants. We clarified our original statement in (Additional File 2, imputation and genetic analyses, line 32-45, page 1). Specifically, from total eMERGE collection of 83717, 9677 unrelated European ancestry participants (1106 cases and 8571 controls) with phenotypic information were selected to be evaluated in this study. This was not clear in our text and we included mode details in additional file 2. After calculating the proportion of variance explained for each PC and examination of scree plot (the elbow of scree plot as shown in Figure S1), the first three PCs were used to account for population stratification. This figure below has been added in Additional file 2-Figure 1S). Please note that in addition to first 3 PCs, the site of genotyping (10 different sites), age, sex and body mass index were also included for a total of 16 covariates. As a result the reported Q-Q plot showed almost no inflation (lambda=1.001) as highlighted in (results, Associations of previously reported SNPs, line 270-271, page 6) and Figure 1b.

We also further investigate this important matter and performed comparison analyses of summary statistics when we include 3PCs, with the total of 16 covariates, versus 10PCs with the total of 23 covariates. Indeed there was no significant differences in the overall summary statistics and results. Two examples are shown in the below table just for this reviewer’s information.

<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>Position</th>
<th>Gene</th>
<th>Minor Allele</th>
<th>MAF</th>
<th>b</th>
<th>OR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs738409</td>
<td>22</td>
<td>44324727</td>
<td>PNPLA3</td>
<td>G</td>
<td>0.23</td>
<td>1.79</td>
<td>1.70 x 10^-20</td>
<td>Using (3 PCs) 16 covariates</td>
</tr>
<tr>
<td>rs738409</td>
<td>22</td>
<td>44324727</td>
<td>PNPLA3</td>
<td>G</td>
<td>0.23</td>
<td>1.78</td>
<td>8.71 x 10^-20</td>
<td>Using (10 PCs) 23 covariates</td>
</tr>
</tbody>
</table>
SNP  CHR  Position  Gene  Minor Allele  MAF  BETA  P
rs5748926  22  17649774  IL17RA  T  0.34  0.91  3.81 x 10^-08  Using 3PCs (16 covariates)
rs5748926  22  17649774  IL17RA  T  0.34  0.90  4.39 x 10^-08  Using 10PCs (23 covariates)

2) In Methods, two cohorts (CCHMC and CHOP) are used to develop and test the NLP algorithm. The NLP algorithm is said to reach a PPV of 95% but it's not clear to which cohort it refers to. Was the algorithm optimized in the training stage by cross validation? Was the testing sample used once or iterated through algorithm optimization, thereby leading to overfitting?

Response:

In eMERGE consortium, we have established a pipeline during the years to construct and validate many different algorithms.

1) First, the primary site constructs the NLP algorithm, extract ICD codes, lab measures and other necessary elements to identify potential cases and controls. Then medical record chart reviews will be performed on a sufficient number of potential cases and controls by an expert physician as a gold standard and after a series of consultation and fine-tuning of the proposed algorithm, the PPV value will be estimated.

2) If this PPV value reach 95% then the algorithm will be validated in secondary site using the same steps.

3) The final NLP algorithm will then be published and become available for public use and available in emerge network web site (https://phekb.org/network-associations/emerge) (ref 17).

We modified and correct the statements in regard to NLP algorithm ((Methods section, Study Participants and Phenotype: line 151 and 156, page4). As described, this algorithm originally was developed in CCHMC and was validated in CHOP in which in both sites PPV value above 95% was obtained.

3) The case-only analysis identifies a genome-wide significant hit between IL17RA and NAS and ZFP90 with fibrosis. These are interesting results but the rather small sample size (n=235) puts a note of caution. It would useful to include the genotype x quantitative trait plots
to evaluate the underlying data distribution. Also, evidence from these new hits on previous GWAS -even nominally- would also give strong support to these new findings.

Response:

Thank you. This is very important issue. Please note that all collected participants have been genotyped in “10 different medical centers” and all were independent (Table 1Sb). For case-control GWAS result we strongly confirmed and replicated previous published genomic effects. For novel quantitative trait such as NAS or fibrosis, we also illustrated a consistent effect of NAS histologic grade for both pediatrics and adults per genotype for PNPLA3 region as shown in figure 3. This satisfies the criteria of independent replication. Please consider that all of the pediatric participants (107 participants) were from CCHMC and the rest (128 participants) were adult who were genotyped in different sites across network. We also reported the estimated power using this sample size as described in (power of study section, line-246, page 6) in which we had >80% power to identify the association for this specific sub-phenotype. As the reviewer suggested, in addition to Figure 3a, we added two additional genotype x quantitative trait plots (figure 3b and 3c) that further support this independent replication with consistent trend of association for Il17RA for NAS (figure 3b ) and ZFP90 for fibrosis (figure 3c) in both pediatrics and adults and per each genotype calls (AA, AB, BB). The Cochran q statistics also has been included (Methods, Statistical analyses, line 225, page 5) that showed no evidence of heterogeneity between adult and pediatric (Cochran q=0.24 and 0.37) respectively. These two additional figures were included in the manuscript (figure 3b and 3c). Finally please also note that some of our results have been independently replicated when we applied PheWAS methodology conditioning on NAFLD status (please see PheWAS section).

4) Pathway based analysis shows a very significant association for IL1 pathway (P=8e-17). Several IL1-family genes co-locate to the same region of the genome. Was the SNP to gene mapping filtered for LD among the associated SNPs (e.g. r^2 > 0.2)? Otherwise this could have inflated the presence of this pathway genes.

Thank you. In all pathway analyses we followed the same rule disregard of region of interest as described in FUMA pipeline. Please note that after the Hypergeometric tests performed (to test if genes of interest are overrepresented in any of the pre-defined gene sets) a secondary multiple test correction is also included and the result shown in our tables all have been adjusted. This particular region on chromosome 2, spans 300KB with 6384 available markers. In order to
address this concern we re-analyzed this region in which we first performed LD pruning with more conservative r2 threshold of 0.2, as the reviewer suggested, and as a result 5391 out of 6384 were excluded. The adjusted result still remained significant p=7.76e-15. However, we agree with the reviewer that due to unusual situation at this locus, we should be cautioned. Therefore, we added a statement in regard to applying a more conservative LD pruning for this specific region in (result section, gene-based and pathway analyses, line 430-433, page 10) as well as Table S6.

Minor comments

In Results, does the top PNPLA3 association differ from adults and pediatric cases? Testing for heterogeneity would be an interesting measure to see differences of patient ascertainment.

Thank you. As previously reported and highlighted in (Result section, Association of previously reported SNPs: line 274, page 7), this effect at PNPLA3 was consistent in both pediatric (p=9.92 x 10-6, OR=1.76 (95%CI=1.37-2.27)) and adult (9.73 x 10-15, OR=1.79 (95%CI=1.55-2.08)) cohorts. We also included the test of heterogeneity statistics as reviewer suggested on this line that indeed shows no evidence of heterogeneity with Cochran’s Q =0.78 and I2=0.

In Results, the replication of previous GWAS hits (Table S3) should inform on the risk allele, OR and Pvalue of the previous study.

Thank you. All of our presented results were consistent in terms of risk alleles, direction of risk alleles and risk allele frequencies with previous publication. We added references as well as OR, P in (Additional File 1, Table S3 (a-b)).

In the results section, very modest evidence is found for epistatic association with PNPLA3 which is most likely to be false positive. Without independent validation, this analysis has modest exploratory value.

Thank you. We agree with the reviewer about the borderline epistatic effect, however the effect was phenotype specific and consistent in both pediatrics and adults. We included the OR of interaction for pediatrics and adults in (Result, Controlling for the main effects at PNPLA3
section, line 316, page 7). Furthermore several proxy markers also support this finding as described that indicate this is not due to potential erroneous behavior of single variant. Since this approach has not been explored previously in the context of PNPLA3 and NAFLD and since the data indicate a biologically relevant finding (i.e., eQTL effects for mitochondrial gene Acyl-CoA Synthetase Medium chain family member 5 (ACSM5) with a key role in oxidative stress), we decided to include it and add a cautionary note in (discussion, line 533, page 12) for the borderline effect as reviewer indicated.

In Results in general, it would benefit from shortening, particularly in the revision of previous genetic hits. Also, methodological details (e.g. pathway analyses) should be moved to the corresponding section or into the supplementary information.

Thank you. We are following the journal policy on this matter that indicates all essential methodology should be in the manuscript. However as reviewer suggested we removed this section plus imputation and PC to “Additional file 2” and this will be further coordinate with the editor.

Reviewer #2 (Stefano Romeo): The work by Namjou et al is a gwas performed in a pediatric and adult population identifying potential novel loci for fatty liver disease. Following are my questions to the authors:

Authors tested approx 7,200,000 SNPs and based on this threshold for genome wide significance should be 7x10^-9, authors should authors consider genome wide significant only those exceeding this p value?

Thank you. We agree with the reviewer that by increasing the number of markers the multiple test correction also need to be adjusted for, however this is under assumption of independent test. As the reviewer is well informed, the LD pattern between the markers is a major issue that reject this independent assumption. In addition, our high-resolution genotyping set is an “imputation-derived marker set” in which imputed markers usually have significant LD with the original genotyping calls according to the ancestry. In fact, in our dataset, if we consider the LD threshold of r2=0.2, 6.9 Mil markers will have to be exclude with only 300,000 “relatively independent markers” remains that gives a Bonferroni correction p-value of 0.05/300K of 1.66E-07. Please note that theoretically, this is still not an absolute independent test. Therefore, we
decided to follow the established standard measure of GWAS threshold of $p=5.00 \times 10^{-8}$ to be consistent with all other publications.

Do authors see an interaction between the HSD17B13 and pnpla3 common SNPs associated with liver disease?

Thank you. Two previous studies showed presence and absence of interaction between PNPLA3 and HSD17B13 respectively (48,49). As the reviewer suggested, we reanalyzed these two regions and found two nominally significant interaction between our best HSD17B13 snp in this study i.e., rs3923441 and rs738409 in PNPLA3 with AST level (adjusted $p=0.01$, Beta-interaction=0.19) as well as ALT level (adjusted $p=0.03$, Beta-interaction=0.16). Of note, these two effects were improved if we restricted only to obese persons ($p=0.002$, Beta-interaction=0.24 for AST and $p=0.02$, Beta-interaction=0.18 for ALT). This pattern indeed is completely consistent with previous report (48). We therefore, included these two new findings on (result, Impact of SNPs on the severity of NAFLD, line 368-374, page 8) and in (discussion, line 604).

Please also note that our best snp is a novel variant that not only has functional eQTL properties but also showed a separate PheWAS effect for abnormal liver enzyme levels ($p=3.74 \times 10^{-6}$, see Additional file 1: Table S6f) in 801 cases and 49860 controls; even after conditioning on NAFLD status with ($p=3.19 \times 10^{-6}$) as described in (PheWAS section, line 490-493, page 11-12). Indeed, this PheWAS result by itself is another presentation of independent replication.

Authors should seek for replication with liver disease in the UK biobank database available

Thank you. While we agree with the reviewers that another replication is desirable, however our main new findings require liver pathology (i.e NAS grading score, fibrosis, inflammation) and currently the available replication cohorts such as UK Biobank do not have the necessary pathology information that would allow replication. Indeed collection of such level of information from eMERGE biobank- repositories took us more than a year. As reported earlier all results presented here were consisted across network despite different genotyping centers or different age group (i.e pediatrics versus adult) with no evidence of heterogeneity. Please also note that we emphasized in discussion the importance of future independent replication in (Discussion, strength and limitation section, line 639 and 647, page15.

The MBOAT7 gene locus has been robustly associated with NAFLD (mancina Gastroenterology 2016) it would be i tested to see the effect of this gene locus in the present cohort.

Thank you. In this study we evaluated all of the previous published variants and only report the positive findings. For this region, although the reported risk allele T for SNP rs641738 at
MBOAT7 were more frequent in our cases (45%) than controls (43%), however it was not significant after controlling for all covariates including BMI (p=0.32). Other sub-phenotypes or other nearby markers also did not show significant findings and we decided not to report this.

In the gwas on liver biopsy how were coded the cases and control groups for fibrosis (yes/no?), authors should provide more details on what F score they use and how they classified fibrosis.

Thank you. We added more information in (Methods section, Study Participants and Phenotype, line 165-174, page 4). We followed the standard method to score NAFLD disease activity in our available cases with liver biopsy as previously described (20). As the reviewer suggested we included a table that further clarifies the NAS score classification system that we applied in this study and has been included in additional file 1 (Table S1c). The NAS is derived from an unweighted sum of scores of liver steatosis (0-3), lobular inflammation (0-3), and hepatocellular ballooning (0-2) and ranging between 0-8. Coexistent fibrosis also has a separate scoring range of (0-4) that consist of no fibrosis (0), Perisinusoidal or periportal (1), portal (2), bridging fibrosis (3) and Cirrhosis (4). As we described, in this study, we performed not only “case-control” GWAS analyses (cases with NAFLD versus healthy controls) but also several “case-only” quantitative analyses to measure disease severity using histologic report and liver enzymes. In the latter, the quantitative scoring values has been used in linear regression model in which we evaluated 235 available cases with NAFLD.

We have responded to the reviewer’ comments to the best of our ability and have made appropriate changes as requested.

The authors have no commercial interests concerning the subject matter or the materials discussed in this manuscript that would constitute a conflict of interest.

Thank you again for your kind consideration.

Sincerely,

Bahram Namjou, MD
Assistant Professor
Center for Autoimmune Genomics and Etiology (CAGE)