Reviewer’s report

Title: GWAS and enrichment analyses of Non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE network

Version: 0 Date: 23 Mar 2019

Reviewer: Antonio Julià

Reviewer's report:

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

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.

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?

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.
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. r2 > 0.2)? Otherwise this could have inflated the presence of this pathway genes.

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.

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

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.

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.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics
Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

Statement on potential review bias
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria”. If your reply is yes to any, please give details below.

I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria
I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal