Reviewer 1 Comments for the Author...

Work by Guo et al. used a gene-based recessive diplotype screening approach to make an exome-wide comparison between controls and the patients with ICD-9 defined hemochromatosis in a cohort housed by the Marshfield Clinic Research Institute. Their analysis directed them to FGF6, which was previously unidentified in iron metabolism, and subsequently validated by the authors to have a functional role through various experiments. The paper was put together coherently and organized well. The writing is good quality. The overall scope of the paper was well stated. The figures and tables were labeled efficiently. The codes used in the analysis were uploaded and made available in Github, which was quite helpful. However, when the computational methodology is concerned, there’re certain details that can’t be overlooked and need to be more clearly stated, especially since the whole premise of the paper relies on how it was carried out. Here’re some questions and comments to the authors to clarify a few important points in the manuscript:

Overall response: We found these points to be thoughtful, insightful, and, most importantly, extremely valuable for aiding us in clarifying the genetic analyses sections. We agree that the original manuscript did not include several important descriptions of the various analyses performed. We believe we have addressed the reviewer’s comments and concerns in the revised manuscript, dramatically improving the transparency of the statistical genetics and design.

1/ The information and the utility of the cohort were very well stated. In concordance with such study data description, it’d be advisable to include statements about the data collection standards (Declaration of Helsinki principles…), informed consent, protocol review etc.

Response: We thank the reviewer for pointing this out. In the first part of the Methods section (Central Wisconsin Hemochromatosis Sample Set), we have included a statement about the adherence to data collection standards:

The study was conducted in accordance with the Declaration of Helsinki. All samples were collected following written informed consent. Marshfield Clinic received a Certificate of Confidentiality from the National Institutes of Health. All investigators using the PMRP samples had obtained Research Ethics and Compliance Training certification through the CITI program. The study protocol was reviewed and approved by the Marshfield Clinic Institutional Review Board (details in **Acknowledgements**).

2/ Individual level filtering was done by matching the ancestry using Principal Component Analysis (PCA) as well as ICD-9 diagnostic codes:

Response: PCA and ICD-9 codes played an important role in selecting our sample set from the larger biobank population. We have further described the use of PCA in our study to exclude genetic background outliers, thereby reducing the likelihood of confounding by population stratification. The entire set of 10,000 exome array-genotyped individuals were subjected to this PCA prior to selecting cases and controls. The >98% of the individuals were of northwestern European ancestry, primarily originating from Bavaria, and the PCA clearly shows a very tightly clustered group of the very large majority of individuals. Given the length requirement of the manuscript, we did not include all of the results from PCA and related dimensionality reduction approaches applied to these genetic data. We have included the following to better explain to the reader the specifics of the PCA-based QC performed:

“Individuals considered genetic background outliers (more than three standard deviations from the centroid of the first two principal components) were excluded from the study. Exhaustive pairwise kinship coefficients were calculated and one individual from pairs of individuals exhibiting third-degree or closer relatedness were removed.”

Both percent transferrin saturation laboratory results and ICD-9 codes (listed in the manuscript) were used as inclusion criteria for cases. Any occurrence of an abnormal percent transferrin saturation laboratory result and/or any instance of any of the ICD-9 codes used to define cases were collectively used as exclusion criteria for the controls.

2.1/ For readers who aren’t familiar with how an ancestry match was done, the authors should explicitly mention which reference populations were used, how the outliers were removed using the PCA and considering which ethnicity. Also, it’s usually advised to spell out any abbreviations used for the first time in the text (i.e. PCA).

Response: We thank the reviewer for their comments regarding clarifying the analysis steps in the manuscript. Importantly, we did not use PCA to match cases and controls. Hence, we did not use reference populations for our study. For this and the majority of studies that we conduct on the Central Wisconsin population, we use PCA to exclude genetic background outliers from inclusion in the case/control sample sets. The reason we do this stems from nature of the Central Wisconsin population where >98% of individuals are of Northwestern European ancestry, primarily from Bavaria. Through extensive research on the genetics of this population, we have found that the very large majority of individuals tightly cluster using the first two or three principal components. We do have a small number of individuals of Hmong, South Asian, and Native American ancestries which are readily identified (the exome genotype array used has a set of ancestry informative markers which amplify the discrimination ability) and can be easily identified and removed to reduce the likelihood of confounding by population stratification. Adjusting by principal components in a regression framework is possible, but we usually prefer not to do this as 1) the adjusted test statistics can have slightly inflated type I and type II error rates as compared to our approach and those effects are accentuated with small counts, 2) computational routines (e.g., permutation approaches) present considerable challenges when applied to adjusted regression analyses, and 3) our primary analysis method of Fisher’s exact test, which is robust to unbalanced case/control numbers, cannot be implemented with covariates. So, we took the approach of simply removing any genetic background outliers through PCA and performing the association analyses on the resulting highly homogeneous sample set. To clarify this, we included information on the removal of genetic background outliers in the Central Wisconsin Hemochromatosis Sample Set section.

2.2/ What’s the extent of genotype-level filtering done on the individuals? These include: genotyping call rates, any relatedness measures (i.e. kinship coefficient) and sex-match between what was reported and found through genotyping.

Response: We thank the reviewer for these insightful comments. Two levels of genotype-level filtering were performed. The initial filtering was performed as part of the AMD consortium (of which the Marshfield samples were a large component) where call rates for each individual had to exceed 0.985, and variants across the AMD consortium had to exceed 0.985 for the genetic marker. Individuals and markers with higher missingness were excluded at this level. Additionally, the first filtering procedure included the removal of variants that departed from Hardy-Weinberg equilibrium at the p<1.0x10-6 level for the AMD consortium sample set. Samples with abnormal sex chromosome findings were also removed. These QC procedures were described previously and cited in the text. As these QC procedures were applied to the entire AMD consortium, we repeated genotype-level and individual-level filtering on the Marshfield sample set. Using Plink, we filtered markers with missing rates exceeding 10% (--geno setting) and we filtered out any individuals exhibiting relatedness with another sample at the proportion IBD=P(IBD=2)+0.5\*P(IBD=1) > 0.20. Additionally, markers departing from Hardy-Weinberg equilibrium within the Marshfield samples at (p<1.0x10-5).

2.3/ Do the final numbers for cases and controls, 18 and 6,896, respectively, reflect these measures?

Response: Yes, the final case and control numbers reflect the individuals left following all of the QC procedures. To clarify in the text, we included the following in the Central Wisconsin Hemochromatosis Sample Set section:

“Of the approximately 10,000 individuals previously subjected to the exome genotyping array and quality control procedures, the phenotype algorithm identified 18 individuals that were selected as hemochromatosis cases. Controls (n=6,896) were individuals without abnormal saturation values and without any instances of hemochromatosis ICD-9 codes.”

3/ While the initial number of variants was stated as being ‘over 500,000’, the subsequent steps of quality control (i.e. genotype call rates, deviation from Hardy-Weinberg Equilibrium (HWE)) were just mentioned in the passing, and how these affected the final distribution of the variants based on frequency (i.e. how many is <= 1% (rare variants), 1-10% (moderately common variants) and >= 10% (common variants)) wasn’t mentioned. Can the authors include this information within text? This would be useful to get a sense on how many rare variants there were left before going into phasing as well as a the final number of variants used in the analysis.

Response: Again, we find the reviewer’s comments very thorough and useful in improving the manuscript. We had 540,101 initial SNPs and 10,861 (2% of total SNPs) variance were filtered out based on HWE (p <10-5). Since the exome array is focused on a more complete interrogation on exonic variants, the frequency of rare alleles was higher than standard GWAS arrays. We found 47.8% probes were rare variants (<1%), 8.1% were moderately common variants (1-10%) and 44.1% probes are common alleles (>10%). To inform the reader, we have included the following statement in the Genotyping section:

“Rare variants (<1% frequency) represented 47.8% of the markers, moderately common variants (1-10% frequency) were 8.1% of the variants, and 44.1% of the variants interrogated were common alleles (>10% frequency).”

[More on the reduction in the number of SNVs from filtering.]

4/ How was the annotation done? How many variants were left after filtering for the functional variants (i.e. missense, nonsense, 5’UTR, 3’UTR, splice sites and “frameshift”)? Were the missense variants annotated for their pathogenicity and also filtered in/out as such?

Response: We applied the ANNOVAR software to annotate the variants. We collected 13 different predictions including phaseConsElements46way, Polyphen2\_HDIV, Polyphen2\_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaLR, M-CAP, CADD and GWAS significant hits. We require at least two “D” preditions to define pathogenic variants. Yes. We define pathogenicity exactly same as the above definition.

5/ Has the unbalance between the number of cases and controls addressed? A lot of times, including rare variants in association studies when there’s such an imbalance would result in increased rates of false positives.

Response: We agree with the reviewer that highly unbalanced designs can present a challenge to genetic analyses and inflate type I error rates. Due to the unbalanced case/control sample sizes in our study, we selected the Fisher’s exact test to perform the association analyses. As the hypergeometric null density of the configuration of the contingency tables holds under all unbalanced designs, there should not be an increased rate of false positives. The detriment to taking this approach is that adjustment for covariates is not feasible with this test. Hence, the statistical analysis is for crude effects. For this study, we viewed the exome scan as a hypothesis generating step which was subsequently validated with the functional experiments. We have included two sentences to the Statistical Tests of Recessive Diplotypes section:

“As the hypergeometric null density holds for all sample sizes, the Fisher’s exact test is robust to unbalance between case and control sample sizes. Simulations have recapitulated this finding showing that the Fisher’s exact test does not inflate type I error rates under unbalanced designs (Ma et al, 2013)”

5.1/ It’d be a good check to look at and compare the mutation burden between the cases and controls using maybe a gene-burden test. Would the findings be similar?

Response: This is an excellent idea. We ran the SKAT test with rvtest (Xuewei Zhan, [Bioinformatics](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4848408/). 2016) on our data for *FGF6* and found the association is still significant, although not as significant as the recessive diplotype test with a Fisher’s exact p-value. The SKAT/rvtest procedure produced a P-value=3.86E-05 and we subsequently ran a permutation routine on this test which produced the permuted p-value=1.0E-04. So, we interpret these results as the association signal can be picked up by the rare variant burden test, but explicitly incorporating the recessive diplotype approach we were found a higher degree of evidence at FGF6 against the null hypothesis of no association with iron overload (p=1.99E-06).

In the Statistical Tests of Recessive Diplotypes section, we included the following sentence:

“To compare the recessive diplotype analysis procedure to a standard rare variant gene-based test, the SKAT test was also applied to the genotype data (REF).”

Within the Results section, we included the SKAT p-values:

“For comparison, the SKAT/rvtest procedure on the *FGF6* genotype data yielded an asymptotic *P=3.86*×10-5 and permuted *P=1.0*×10-4.”

6/ The most important part of the analysis revolves around the details surrounding the haplotype phasing step in the presence of both common and rare variants, which were not sufficiently explained. Here, the purpose of phasing is to ascertain compound heterozygosity from the genotyping data. Given a string of heterozygous calls, we actually don’t know which are in phase and anti-phase. The ultimate proof would be to experimentally show that the adjacent alleles are in phase by long-range PCR amplicon sequencing or single molecule long-read sequencing. Statistical phasing only works well for common variants for which one can tap into common haplotypes in a reference database. However, for rare and ultra-rare variants, statistical phasing is not as reliable. It is not clear in the manuscript which database or reference was used for phasing and how strong the statistical support was. A great deal depends on how large the reference sample is (that has been typed on the same platform) and whether the rare variants had many observations to do credible phasing. If phasing is not reliable, then the statistical uncertainty or the so-called switching error ought to be documented.

The argument about how in a string of k variable sites, how it is unlikely all rare variants are on one side, needs to be examined in a case-by-case basis. Because if the cases have more suspected heterozygosity because they just have more rare variants; then the hypothesis of rare variant burden difference between cases and controls need to be revisited. Therefore, it would be useful to see a lot more details and summary information on how the phasing was done and how its performance was.

Response: We agree with the reviewer that the phasing is an exceedingly important step in conducting the recessive diplotype scan. We also agree that rare variants are often challenging to phase correctly. We agree that additional details on the phasing approach will significantly improve the manuscript. We would like to make four points on this topic: 1) the use of our highly homogeneous population with recent shared ancestry improves phasing accuracy, 2) the use of very large sample sizes (n=10,000) also improves phasing accuracy, 3) we employed the Beagle software for phasing which is highly regarded in the statistical genetics community, and 4) the original Beagle paper showed that for medium to large sample sizes, the phasing error rate was between 0.77-0.94% and a recent study (Choi et al, 2018) has shown that Beagle produced switch error rates of 1.525% for the 1000 Genomes Project data and 0.488% for the Haplotype Reference Consortium—given that these are sequencing-based data sets, the number and frequency of rare variants substantially exceeds those in our study (exome genotype array). We undoubtedly have phasing errors in our sample set, but these studies strongly suggest that the error rates are small and given that the phasing was performed in our study agnostic to case/control status, the rate of errors is likely evenly distributed between cases and controls; this may reduce the power to detect disease-predisposing effects from compound heterozygotes, but may not substantially inflate false positive rates. We view the recessive diplotype scan as an easily implemented procedure that can be applied to existing data sets which can generate candidate genes for further functional interrogation, subsequently validating the plausible role in disease susceptibility. In our study, we believe that the functional studies indicate that *FGF6* is not a false positive.

We agree that the k variant argument needs to be applied on a sample-by-sample basis. Our intention was to present an argument that for some samples with a large number of rare, putatively pathogenic variants at a certain gene, one can calculate the lower bound for the probability that the individual is indeed in the recessive diplotype class, regardless of the phasing procedure. We believe that this argument may be useful in selected circumstances when others conduct a recessive diplotype scan.

To clarify the phasing procedure, we …

7/ Comparative genomic analysis of the iron metabolism genes can be moved to the discussion in a supportive fashion rather than part of the results as it may not be completely a validation point or proof for the function of FGF6.

Response: We agree with the reviewer. We have moved the comparative genomic analysis to the Discussion section and the corresponding Figure is re-assigned to be supplementary Figure xx.

Reviewer 2 Comments for the Author...

By using a gene-based compound heterozigosity test on 10,000 DNAs the authors identified HFE and FGF6 as susceptibility genes of iron overload. The finding is interesting and so is the comparative genomic analysis showing that FGF6 evolved together with major genes of iron metabolism. However, the cellular

studies to support the role of FGF6 in cellular iron homeostasis seem preliminary and not fully convincing.

1- The studies of cellular iron supplementation should be done more carefully. The iron concentration and time of incubation should be stated. Moreover the quantification of cellular iron uptake by counting the Perl’s positive cells is not straightforward. Perl’s stains mainly hemosiderin iron which is only a

fraction of total iron and that is deposited only when iron is in excess. Moreover, I do not see the blue cells in the strong red background in fig 3 and 4. More direct methods would be more convincing, such as measurement of total cellular iron, or of ferritin protein, ferritin-iron or also transferrin receptor mRNA.

Answer: Thanks for the question and suggestion. In our study, iron supplementation changed to ferric ammonium citrate (FAC) in the presence of ascorbate. The concentration of ascorbate and FAC was 500 μM and 10 μM which was suggested by Sinead Healy et al (Prog Neurobiol. 2017 Nov;158:1-14) and the time of incubation was 48h in Ferrozine assay, Perl’s stains, Western blot. In order to quantify cellular iron, we use the ferrozine assay (Anal Biochem. 2004 Aug 15;331(2):370-5.) and the protein level of Ferritin by western blot. In addition, the intracellular accumulation after FAC treatment was also confirmed by Perl’s staining (supplementary figure). The conclusion was in accord with previous research.

2- The differential uptake of Fe(II) and Fe(III) cannot be used as an evidence of hepcidin-dependent iron uptake. Fe(III) at neutral pH readily forms polynuclear insoluble complexes the absorption of which is not studied. Fe(II) is supposed to be taken up mainly by ZIP14, which is unrelated to hepcidin activity. Most studies of iron supplementation use ferric ammonium citrate in the presence of ascorbate to maintain it in a mononuclear and soluble form.

Answer: We completely agree with reviewer’s suggestion. FAC and ascorbate applied to iron incubation experiments.

3- The finding that FGF6 overexpression induces hepcidin mRNA is rather convincing, less clear is the effect of the three mutants. The claim that M2 (D174V) differs from M1 (E172X) and M3 (R188Q) is supported only by fig 4B and 4D, and not by fig 4E-G.

Answer: Thanks for the comments. M2 (D174V) differs from M1 (E172X) and M3 (R188Q) was supported by Ferrozine assay, Western blot, Perl’s stains.

4- A list of abbreviations would help the reader.

Answer: Thanks for the suggestion. We added a list of abbreviations to help the reader understand the paper.