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. Changes to the manuscript are denoted in red font.

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. 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**).

The Acknowledgements section also includes additional details on the IRB approval. We have added the following statement to this section:

Marshfield Clinic received a Certificate of Confidentiality from the National Institutes of Health.

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 was 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 in the **Methods** section (within the **Central Wisconsin Hemochromatosis Sample Set** subsection) to better explain to the reader the specifics of the PCA-based QC performed:

“To reduce confounding by population stratification, a Principal Components Analysis (PCA) on the exome genotyping data was implemented using all samples, blinded to disease status. Individuals considered genetic background outliers (more than three standard deviations from the centroid of the first two principal components) were excluded from the study. Following the removal of outliers, the resulting set of individuals was highly homogeneous based on the first three principal components. Exhaustive pairwise kinship coefficients were calculated and one individual from pairs of individuals exhibiting three-degree of 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. This information is provided in the **Methods** section under the **Central Wisconsin Hemochromatosis Sample Set** subsection.

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 useful 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). In the **Genotyping** subsection of the **Methods** section, we have clarified these points with

“The genotyping quality control measures were previously described (call rates for each variant or individual >0.985).34 Variants exhibiting departure from Hardy-Weinberg equilibrium (p<1x10-6) were excluded from subsequent analyses.”

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** subsection:

“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%). 413,701 SNPs remained after removing 126,400 monomorphic sites. Further, we compared the minor allele frequency (MAF) distribution with the final SNP set after the filtering for functional variants and found that the MAF distributions were not significantly different (p=0.22). These distributions are now displayed in **Supplementary Figure 2** and **Supplementary Table 1**.

To better inform the reader, we have included the following statement in the **Genotyping** subsection:

“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).”

Additionally, we have included the following sentence in the **Genotyping** subsection:

“Following QC procedures, 413,701 variants remained for analysis. The site frequency spectrum of the resulting variants is displayed in **Supplemental Figure 2**.”

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 started the filtering process with variants that were previously GWAS significant, missense, nonsense, 3’UTR, 5’UTR or occurring in splice site regions. After obtaining this set, we applied the ANNOVAR software to annotate the variants. We collected 12 different predictions including phaseConsElements46way, Polyphen2\_HDIV, Polyphen2\_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaLR, M-CAP, and CADD. We required that at least two “D” predictions to define pathogenic variants. We defined pathogenicity using the above criteria and variants meeting this set of criteria were then deemed putative functional and included in the gene-based analyses. To clarify this in the text, we have included in the following in the **Determination of Putative Functional Variants** subsection:

“Additionally, on the resulting set of variants, annotation was performed using the ANNOVAR software.38 Only variants that were also annotated as pathogenic by at least two ANNOVAR predictions were included in the scan. Following filtering for putative functional variants, 129,556 SNPs remained for use in our gene-based recessive diplotype scan. The putative functional variants of *FGF6* are shown in Supplemental Table 2.”

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. That said, to investigate the effect of sex on the *FGF6* association (sex plays a large role in the prevalence of iron overload), we partitioned the data by sex, then calculated a Mantel-Haenszel test of homogeneity to determine if there was statistical evidence for sex-specific differences in effect at *FGF6*. The OR within male and female categories was similar and not significant (p=0.728).

We have included two sentences to the **Statistical Tests of Recessive Diplotypes** subsection:

“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.39”

We have also included the following to address the investigation of differences in sex-specific effects in the **Statistical Tests of Recessive Diplotypes** subsection:

“Additionally, to investigate the sex-specific effects, the Haldane OR was calculated separately for female and male strata. Lastly, the Mantel-Haenszel test of homogeneity was calculated to determine the level of statistical evidence for sex-specific differences in effects.42

Accordingly, we report the findings of the test for sex differences in the **Gene-based compound heterozygosity scan identifies a novel hemochromatosis susceptibility gene** subsection of the **Results** section:

“Additionally, no statistical evidence of effect differences between females and males for the *FGF6* data (Mantel-Haenszel test of homogeneity *P*=0.728).”

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 as it places our analysis within the context of established methods. 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=1.0x10-4. We interpret these results as the association signal can be picked up by the rare variant burden test (although not significant exome-wide), 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=1.99x10-6) (significant exome-wide).

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 RVTESTS software40 which implements the sequence kernel association test was also applied to the genotype data.41 “

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. Notably, the recessive diplotype scan result exceeded exome-wide significance, whereas the rare variant test did not.”

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. That said, we believe this argument is outside the scope of the present study (we did not use this argument in conducting the study) and needs to be examined more extensively. Therefore, we have removed the section discussing the k variant argument.

To address these points in the text (**Haplotype Phasing** subsection) and clarify our procedures to readers, we clarified that the entire set of 10,000 exome-genotyped samples were used in the phasing to increase phasing accuracy:

“Using all 10,000 exome-genotyped samples from the PMRP, the software package Beagle was applied to infer phased haplotypes from the unphased genotype data using a localized haplotype-cluster model algorithm.36

We also provided information on how these calculations were performed:

“The calculations were performed on a high performance computing cluster housed at the Marshfield Clinic. As the subsequent analyses were gene-based and the genotyping data was concentrated on exonic variants, each gene in the exome was phased separately using this approach.

To address the issue of phasing accuracy, we included the following in the **Haplotype Phasing** subsection. Importantly, we cite a recent study showing that Beagle has mild to moderate error rates even when applied to sequence data, which will have a higher frequency of rare variants than the exome genotyping array used in our study:

“Although rare variants can present difficulties in phasing, the use of a large sample size from a highly homogeneous population aids in mitigating the error rate. Notably, Beagle has been shown to have error rates in phasing between 0.77-0.94% for medium (n=1,000) to large (n=5,000) sample sizes using a 500K GWAS array.36 Recently, the switch error rate was calculated for the Beagle software applied to two sequencing datasets. Beagle attained a switch error rate of 1.525% and 0.488% for the 1000 Genomes Project and Haplotype Reference Consortium, respectively.37

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 4**.

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.

Response: We would like to thank the reviewer for the knowledgeable insights and excellent suggestions that have served to dramatically improve the manuscript. Iron supplementation was changed to ferric ammonium citrate (FAC) in/with the presence of ascorbate. The concentration of FAC and ascorbate was 10 μM and 500 μM which was suggested by Healy et al (Prog Neurobiol. 2017 Nov;158:1-14). The time incubation was stated in detailed in different experiments. In order to quantify cellular iron following the suggestion, we used the ferrozine assay and the protein level of Ferritin by western blot. The ferrozine assay is a sensitive, simple, reproducible and accessible method that is widely used for detecting iron content in cultured cells, comparable results were obtained between the ferrozine assay and Atomic absorption spectroscopy (AAS) which is the most accepted analytical technique for iron analyses (Prog Neurobiol. 2017 Nov;158:1-14). Within cells, ferritin is the iron-storage protein which can quantify iron. The time of iron incubation had a smaller difference in various experiments. To measure total iron content after FGF-6 active protein treatment, cells were seeded in 12-well plate, and then treatment with FAC and different amount of FGF-6 protein (0, 2.5, 25 and 250ng/ml), after 48h, the cells were harvested and total iron contents were measured with Ferrozine assay. To measure total iron concentration after *FGF6* mRNA treatment, the cells were seeded in 24-well plate and transfected with *FGF6* wildtype and three mutant variants, after 6h, the culture media was completed removed. Fresh media with 10% FBS and FAC were added, after 48h, the cells were harvested to perform ferrozine assay, Perls’ stain and western blot. In RT-PCR and quantitative RT-PCR analysis, cells were treated with FGF-6 active protein or transfected with *FGF6* plasmids for 24h. Surprisingly, we found that FGF-6 has more significant results in decreasing total iron content and ferritin level. M1 and M3 had larger effects from wildtype compared to M2. We view these results as providing evidence that FGF-6 is associated with iron metabolism, and that M1 and M3 may confer more profound loss-of-function effects than M2 which is expected since M1 is a stopgain variants (Figure 4G and 4H) which have most serious effect to protein function.

In the **Methods** section within the **Cell culture, Reagents, and Protein treatment** subsection, we added the following:

“Colon cancer cell lines (HCT8 and HCT116), a kidney cancer cell line (786-O), a liver cancer cell line (HepG2) and a fibroblast cell line (HFF-1) were cultured in DMEM medium supplemented with 10% FBS at 37°C in a 5% CO2 humidified incubator. To investigate the change in iron uptake under different protein treatments or plasmid transfection, the Ferric Ammonium Citrate (FAC) cell culture method was employed. Cells were cultured in normal DMEM and FBS medium with the presence of 1uM FAC and 50uM ascorbate for 48h during detection of cellular iron concentration. Cells cultured in normal medium exhibited very low iron concentrations. Total intracellular iron concentration in cells were considerably higher than cells Recombinant human FGF-6 protein (Active) (ab219122), anti-Ferritin (ab75973) were purchased from Abcam, Flag tag antibody (20543-1-AP) was purchased from Proteintech Group, anti-GAPDH antibody was purchased from Shanghai Yeasen Biotechnology. 1mM Ferric ammonium citrate and 50mM ascorbate were dissolved in distilled water. NaOH, HCl, KMnO4, ferrozine, neocuproine, ammonium acetate, ascorbic acid and FeCl3 were purchased from Beijing Oka Biological Technology. Plasmid with raw FGF6 sequence was purchased from PPL-Shanghai Co., Ltd (Shanghai, China) which was constructed in an N-Terminal p3XFLAG-CMV vector, whereas three different FGF6 mutations (E127X, D174V, R188Q) were synthesized with overlapping-PCR.

**Quantification of Iron content by ferrozine assay**

Total intracellular iron content was measured by the ferrozine assay.45 Cells were cultured in 12 or 24-well plate for 48h and washed three times with cold PBS. After lysed 2h with 50mM NaOH, 100μL of cell lysates were mixed with 10mM HCl, and 100μL of the iron-releasing reagent (a freshly mixed solution of equal volumes of 1.4M HCl and 4.5% (w/v) KMnO4 in H2O). The mixtures were incubated for 2h and 30μL iron detection reagent (6.5 mM ferrozine, 6.5 mM neocuproine, 2.5M ammonium acetate, and 1M ascorbic acid) was added, after 30min incubation, 280μL of solution was added to a 96-well plate and read 550nm on a microplate reader. In addition, FeCl3 (0-100μM) was used as iron standards and protein quantification was determined by a Lowry protein assay.

**Western blot**

Cell lysates were harvested when incubated iron for 48h, then equal amounts of protein from every sample were subjected to 12% SDS-PAGE gels electrophoresis and then transferred to PVDF membranes. After blocked with 5% BSA, the membranes were incubated with GAPDH (1:10000), Ferritin (1:1000) and Flag (1:2000) at 4°C overnight. Then, membranes were washed three times with TBST, incubated with anti-rabbit or anti-mouse secondary antibody. The bands were visualized using Image QuantTL software.

**Perls’ staining**

Cells were washed with phosphate-buffered saline (PBS) three times, fixed with 4% glutaraldehyde for 10 min, and incubated at 37°C for 60 min with 2 ml Prussian blue solution comprising equal volumes of 2% hydrochloric acid aqueous solution and 2% potassium ferrocyanide (II) trihydrate. After the cells were stained with 0.5% neutral red for 3 min, iron staining was visualized by Nikon microscope. Iron positive high positive staining cells divided by total cell number used to evaluate the iron deposition levels.”

In the **Results** section, **Figure 3** was changed to show iron content in various cells (HepG2, 786-O, HCT-116, HCT-8 and HFF-1) after FGF-6 active protein and FAC incubation. In **Figure 4**, iron content and ferritin protein level in cells (HepG2 and HCT-116) after *FGF6* mRNA transfection and FAC incubation was added. Perls’ stain results were removed to **Supplementary Figure 5**. In **Figure 5**, IHC of ferritin in SSc skin tissues and liver cancer tissues was added, perls’ stain in SSc skin tissues and liver cancer tissues were removed to **Supplementary Figure 7**.

Accordingly, within the **FGF-6 modulation of hepcidin expression and iron uptake** subsection of the **Results**, we have included the following:

“To investigate the potential mechanism linking FGF-6 to iron metabolism, the effects of FGF6 on iron uptake and the expression of iron-metabolism genes in HepG2, HCT8, HCT116, 786-O and HFF1 cells were evaluated. Using cells cultured cells and a ferrozine assay to detect iron, total intracellular iron concentration was significantly decreased in HepG2, 786-O, HCT8, HCT116 and HFF-1 cells when treating with active FGF-6 protein in a dose-dependent manner (**Figure 3A-D and Fig. S5**). Testing the effect of FGF6 on the expression of a set of genes involved in iron metabolism (*HAMP*, *HDAC2*, *HMOX1*, *TFRC*, and *HEPH*), HepG2 cells were subjected to treatment from control or FGF-6 protein and *FGF6* mRNA or control and mRNA expression relative to *GAPDH* was measured in the five iron metabolism genes. RT-PCR analysis revealed that *HAMP* and *HDAC2* mRNA levels were significantly increased after the FGF-6 active protein introduction in HepG2 cells compared to treatment with PBS as control (**Figure 4A**). *FGF6* plasmid transfection significantly increased *HAMP*, *HDAC2*, and *HMOX1* levels, whereas *TFRC* levels significantly decreased in HepG2 compared to a vector without *FGF6* (**Figure 4B**). *HEPH* expression did not change with either *FGF6* plasmid or FGF-6 protein, suggesting the effect of FGF-6 may be independent of *HEPH* (**Figures 4A, 4B**).”

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.

Response: We completely agree and take reviewer’s suggestion. Instead of focusing on Fe(II) and Fe(III), FAC and ascorbate was applied to all iron incubation experiments. The concentration of ascorbate and FAC was 500 μM and 10 μM which was suggested by Healy et al (Prog Neurobiol. 2017 Nov;158:1-14). We also agree that the conclusion of a hepcidin-dependent iron uptake was not correct and have removed the claim from the manuscript.

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: This is an excellent point. Within the HCT-8 cell line, we did not see a significant difference in hepcidin mRNA expression with the introduction of plasmids containing M1, M2 or M3. That said, in other cell lines (786-O, A498 and HepG2), we did observe a significant difference in hepcidin mRNA expression between wildtype and the mutant plasmids. Hence, we suspect that the differences may be cell line-specific. For the revision, we have focused on HepG2 and HCT-116 (new cell line) to perform additional experiments on the mutants including a Ferrozine assay, Western blot and Perls’ stain. So, the revised **Figure 4** presents results for all of these assays on the HepG2 and HCT-116 cell lines to provide a more complete picture of the effect of the mutants on these two cell lines. We have moved the 786-O, A498 and HCT-8 results of mutant plasmids on hepcidin mRNA expression to **Supplemental Figure 11**. While the results paint a fairly complicated picture of the M1, M2 and M3 effect on hepcidin mRNA expression, FGF-6 expression and iron concentration in different cell lines, we believe these experiments provide preliminary data to motivate additional research on the function of these mutant alleles. Accordingly, we have tempered our statements on the effect of the mutant alleles. To summarize, we observed:

1. A significant reduction in hepcidin mRNA expression by M1 compared to WT in 786-O, A498, HepG2, and HCT-116, but not HCT-8.
2. A significant reduction in hepcidin mRNA expression by M2 compared to WT in HepG2, but not 786-O, A498, HCT-8, or HCT-116.
3. A significant reduction in hepcidin mRNA expression by M3 compared to WT in 786-O, A498, HepG2, and HCT-116, but not HCT-8 (same pattern as M1).
4. A significant increase in iron concentration by M1 compared to WT in both HepG2 and HCT-116.
5. A significant increase in iron concentration over WT was not observed for M2 in either HepG2 or HCT-116.
6. A significant increase in iron concentration by M3 compared to WT in both HepG2 and HCT-116 (same pattern as M1).

In the results section, we added a subsection which reads:

**Evaluation of FGF6 variants on HAMP expression and iron concentration compared to wildtype *FGF6***

In order to investigate the effects of the *FGF6* alleles on FGF-6 function, we transfected plasmids carrying either the wildtype *FGF6* or variant *FGF6* with each of the three point mutations described above (**Figure 2 and Fig S6**). The M1 (E172X) and M3 (R188Q) variants exhibited a significant downregulation of *HAMP* mRNA compared to WT in HepG2 cells (**Figure 4C**), HCT-116 cells (**Figure 4D**), 786-O (**Fig S11**), A498 (**Fig S11**), but not HCT-8 cells (**Fig S11**). Evaluating the effect of M2 (D174V) on *HAMP* expression compared to WT only yielded a significant reduction in HepG2 (**Figure 4C**), but not in any of the other cell lines (**Figure 4D**, **Fig S11**). Further, we noted *HAMP* mRNA levels in M1 and M3 transfections were comparable to control levels, which illustrated a strong attenuation of *FGF-6* function for M1 and M3 variants (**Figure 4C-D, Figure S7A**). Examining the impact of specific variants on intracellular iron concentration in HepG2 and HCT-116 cells, M1 and M3 produced significantly elevated iron deposition **(Figure 4E, 4F, Fig. S7B)** and ferritin expression (**Figure 4G-H, Fig. S7C-D**) indicating a deficiency in M1/M3 *FGF6*-mediated iron homeostasis. In addition, the intracellular iron accumulation pattern was confirmed by IHC using Perls’ stain **(Figure S5)**. In contrast, M2 did not produce a significant departure from WT in iron concentration and ferritin expression in HepG2 and HCT-116 (**Figure 4**). Further, TFRC expression was significantly upregulated in the presence of M3 compared to WT (**Figure 4C**). The functions mentioned above were also validated in HFF-1 (**Figure S8**).

4- A list of abbreviations would help the reader.

Response: We appreciate the suggestion. We added a list of abbreviations to aid the reader:

**Abbreviations**

GWAS: Genome-wide Association Studies

EHR: Electronic health record

PMRP: Personalized Medicine Research Project

PCA: Principal Components Analysis

AOD: Average optical density

FAC: Ferric ammonium citrate

HCT8: Human ileocecal colorectal adenocarcinoma cell line

HCT116: Human colon carcinoma cell line

786-O: Human kidney adenocarcinoma cell line

HepG2: Human liver hepatocellular carcinoma cell line

HFF-1: Human skin fibroblast cell line

HBS: Heparin binding sites

HAMP: Hepcidin antimicrobial peptide

HEPH: Hephaestin

*FGF6*: Fibroblast Growth Factor 6 gene

FGF-6: Fibroblast Growth Factor 6 protein

M1: E172X *FGF6* variant (NM\_020996.2:c.514G>T)

M2: D174V *FGF6* variant (NM\_020996.2:c.521A>T)

M3: R188Q *FGF6* variant (NM\_020996.2:c.563G>A)

SSc: Systemic sclerosis