## A Gene-Based, Recessive Diplotype Exome Scan Discovers *FGF6* as a Novel Iron Metabolism Gene

Shicheng Guo1#, Shuai Jiang2#, Narendranath Epperla3, Yanyun Ma2, Mehdi Maadooliat1,4, Zhan Ye5, Brent Olson5, Minghua Wang6, Terrie Kitchner1, Jeffrey Joyce1, Robert Strenn5, Joseph J. Mazza7, Jennifer K. Meece8, Wenyu Wu9, Li Jin2, Judith A. Smith10, Jiucun Wang2,11,\*, Steven J. Schrodi1,12,\*

1Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA

2State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

3Division of Hematology, The Ohio State University, Columbus, OH, USA

4Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI, USA

5Biomedical Informatics Research Center, Marshfield Clinic Research Institute, Marshfield, WI, USA

6Department of Biochemistry and Molecular Biology, Medical College, Soochow University, Suzhou, China

7Clinical Research Center, Marshfield Clinic Research Institute, Marshfield, WI, USA

8Integrated Research and Development Laboratory, Marshfield Clinic Research Institute, Marshfield, WI, USA

9Department of Dermatology, Huashan Hospital, Fudan University, Shanghai, China

10Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA

11Institute of Rheumatology, Immunology and Allergy, Fudan University, Shanghai, China

12Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA

#These authors contributed equally to this work

\*Correspondence:

Jiucun Wang, Ph.D.

School of Life Sciences

Fudan University

Shanghai, China

Email: [jcwang@fudan.edu.cn](mailto:jcwang@fudan.edu.cn)

Steven J. Schrodi, Ph.D.

Center for Human Genetics

Marshfield Clinic Research Institute

1000 N Oak Ave--MLR

Marshfield, WI 54449

Tel: 715-221-6443

Email: [Schrodi@wisc.edu](mailto:Schrodi@wisc.edu)

## Abstract

Standard analyses applied to genome-wide association data are well-designed to detect additive effects of moderate strength. However, the power for standard GWAS analyses to identify effects from recessive diplotypes is not typically high. We proposed a gene-based compound heterozygosity test to reveal additional genes underlying complex disease heritability. With this approach, a strong association signal was identified between the fibroblast growth factor-encoding gene, *FGF6,* and hemochromatosis, a common iron metabolism disorder, in the central Wisconsin population. Functional validation showed FGF-6 regulates iron homeostasis and induces transcriptional regulation of hepcidin. Moreover, specific identified *FGF6* variants differentially impact iron metabolism. In addition, FGF6 downregulation correlated with iron metabolism dysfunction in systemic sclerosis and cancer cells. Using the recessive diplotype approach revealed novel susceptibility genes and has extended our understanding of the mechanisms involved in iron metabolism.

Genome-wide Association Studies (GWAS) are well-designed to detect additive effects of modest effect sizes. We hypothesized that gene-based tests sensitive to recessive diplotypes—including recessive single site effects and compound heterozygosity—may reveal additional genes underlying complex diseases. Carrying variants conferring a compromised function on both homologous chromosomes is likely to impact molecular physiological states. Deep sequencing studies have conclusively shown a vast reservoir of rare variants segregating in human populations.[1](#_ENREF_1) Rare variants in functional categories (e.g., missense, regulatory motifs) may generate pathogenic effects through recessively-acting diplotypes, and such effects are apt to remain concealed from standard GWAS analyses. Simple power calculations show that recessive diplotype inheritance produces signals that are difficult for standard GWAS methods to discover (**Supplementary Figure 1**). Recently, a thorough investigation of compound heterozygosity disease models has provided ample evidence that geneticists should consider these effects.2 These models also enjoy a high degree of biological plausibility, particularly if the alleles confer compromised protein function.[2-5](#_ENREF_2) Recessive diplotype modes of inheritance are well-established in Mendelian diseases, such as cystic fibrosis[6](#_ENREF_6), mevalonic aciduria[7](#_ENREF_7), beta-thalassemia[8](#_ENREF_8) and Niemann-Pick disease.[9](#_ENREF_9) Although not systematically examined in population-based studies, there is a sizable repository of genes underlying complex diseases with recessive, loss-of-function effects.[9-14](#_ENREF_9) Hence, we posited that an exome-wide, gene-based screen of recessive diplotypes using putative functional variants in both oligogenic and complex diseases may expand our knowledge of disease genes.

Iron metabolism disorders, including adult hereditary hemochromatosis, collectively are common conditions with considerable public health implications.[15](#_ENREF_15),[16](#_ENREF_16) Importantly, the hepatic hormone hepcidin is a key regulator of iron homeostasis by controlling iron flux from enterocytes and macrophages to plasma through degradation of the cellular iron exporter ferroportin. To investigate the inheritance of hemochromatosis, several segregation analyses were initially conducted, concluding that a recessive mode of inheritance is highly plausible.[17](#_ENREF_17),[18](#_ENREF_18) A number of human studies have investigated the genetics of iron overload, revealing several critically important genes. Notably, *HFE*, encoding the membrane-bound hereditary hemochromatosis protein, was mapped two decades ago through family-based linkage[19-22](#_ENREF_19) and association approaches.[23](#_ENREF_23),[24](#_ENREF_24) Additional studies have definitively placed the missense polymorphism C282Y in *HFE* as the major susceptibility factor in adult-onset, type 1 hereditary hemochromatosis.[25](#_ENREF_25),[26](#_ENREF_26) Additional genes have been identified through pathway-based genetic association studies and GWAS, including *BMP2, BMP4, HJV, TF, TMPRSS6, NAT2, FADS2*, and *TFR2*.[27-29](#_ENREF_27)

#### Gene-based compound heterozygosity identified novel hemochromatosis susceptibility gene

To discover novel iron overload-predisposing genes, we conducted a gene-based scan for recessive diplotypes composed of putative functional alleles across the exome using biobanked samples linked to electronic medical records obtained from a rural, genetically-homogeneous population in central Wisconsin. Of the 10,000 samples evaluated, our transferrin saturation and diagnostic code-based phenotype algorithm identified 18 case individuals and 6896 controls. We estimated gametic phase on all individuals and restricted our analyses of diplotypes to putative functional variants. Our recessive diplotype scan identified two exome-wide significant genes (**Figure 1, Table 1 and Supplementary Figure 2**), *HFE* (*P=*1.29×10-8; OR=28.7) and *FGF6 (P=1.99*×10-6; OR=22.8*)*. This motivated our investigation of FGF-6 function and the impact of specific *FGF6* variants on iron metabolism.

#### Comparative genomic analysis reveals FGF6 evolved synchronously with iron metabolism genes

To explore the involvement of *FGF6* in iron metabolism, we conducted the evolutionary analysis of *FGF6* and known iron metabolism genes including *FGFR1, TFRC, FTH1, IREB1, TF, HMOX1, ACO2* and *HAMP* (encoding hepcidin). The appearance of iron metabolism genes can be separated into two stages. *TF* and *HMOX1*, which are found in animals from *C. elegans* to *H. sapiens*, indicates an origin in early Bilateria evolution (~635 Mya). *FGF6*, *FGFR1*, *ACO2* and *HAMP* can be found from *D. rerio* to *H. sapiens*, but are not present in *C. elegans* and *Drosophila*, indicating emergence in early Vertebrata (~485 Mya). The co-appearance of these genes suggests possible co-regulatory functions (**Figure 2A**). Investigating FGF-6 protein-protein interactions, we found evidence for FGF-6 interactions with FGFR1, MAPK1/3, INS, FN1 and involvement in the iron metabolism subnetwork involving TF, HFE, HAMP and SLC40A1 (**Figure 2B**). FGF-6, also known as Heparin Secretory-Transforming Protein 2 or Heparin-Binding Growth Factor 6 has multiple heparin binding sites (HBS). Three known nonsynomymous variants located in the heparin binding sites (R188Q) or flanking sites (D174V and E172X) were speculated to be important for the FGF-6 function. Further, D174V and E172X are located in the regions between FGFR-binding region (FGFR-BR-3) and HBS-1 (**Figure 2C**). Hence, we studied these three variants in functional studies to further investigate the involvement of FGF-6 in iron metabolism.

#### FGF-6 associated with modulatation of hepcidin expression and hepcidin-dependent iron uptake

To investigate the potential mechanism linking FGF-6 and iron metabolism, we evaluated iron uptake and the expression of iron-metabolism genes in human liver carcinoma HepG2 cells. Staining indicated that ferrous (Fe2+), ferric (Fe3+),and a mixture of Fe2+ and Fe3+ forms can significantly increase intracellular iron deposition (**Figure 3A**). We then co-treated the cells with Fe2+，Fe3+ and FGF-6 active protein, which resulted in a significant FGF-6-mediated decrease in uptake of Fe2+ but not Fe3+ (**Figure 3B**). HepG2 and renal cell adenocarcinoma (786-O) cells were then transfected with the plasmids carrying either the wildtype *FGF6* or variant *FGF6* with each of the three point mutations described above (**Figure 2C**) and evaluated for iron deposition. Two variants, M1 (E172X) and M3 (R188Q), had elevated iron deposition **(Figure 4A and Figure 4B),** indicating a deficiency in variant FGF-6-mediated iron uptake-suppression. RT-PCR analysis revealed that *HAMP* and *HDAC2* mRNA levels were significantly increased after the FGF-6 active protein treatment, with no significant change in *HMOX*, *TFRC* and *HEPH* (**Figure 4C**). *FGF6* plasmid transfection in HepG2, A-498 (kidney carcinoma cells), 786-O increased *HAMP*, *HDAC2*, and *HMOX* levels, whereas *TFRC* levels significantly decreased (**Figures 4D-G**). *HEPH* expression did not change, suggesting the effect of FGF-6 is independent of *HEPH*. 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 4D-G**). The M2 variant (D174V) significantly differed from wildtype in *HAMP* induction only in HepG2 cells.

#### Altered FGF6 gene expression in systemic sclerosis and cancer

We hypothesized that FGF-6 might be involved in human autoimmune diseasesandcancers since abnormal iron metabolism has been reported in numerous studies.[30-33](#_ENREF_30) More specifically, decreased hepcidin has been implicated in the anemia of chronic disease which frequently accompanies these systemic inflammatory states. To explore the relationship between *FGF6* expression and iron deposition in autoimmune tissues, *FGF6* expression and iron deposition in the skin lesions from systemic sclerosis patients (SSc) and healthy controls were examined. Decreased FGF-6 protein level and elevated iron deposition were detected in SSc skin tissue, especially in the epidermis (**Figure 5A-B**). Negatively correlating with intracellular iron, FGF-6 expression was significantly decreased in SSc skin (**Figure 5E**). In investigating liver cancer and adjacent normal tissue for iron deposition and FGF-6 expression, non-metastatic cancer lesion tissues had increased iron deposition and lower FGF-6 expression compared to controls (**Figure 5C, 5D and 5F**). However, increased FGF-6expression was observed in metastatic liver carcinoma tissue (**Supplementary Figure 5A-B**), suggesting that FGF-6 plays different roles in oncogenesis and metastasis, analogous to TGF-.[34](#_ENREF_34),[35](#_ENREF_35)

## Discussion

Iron homeostasis results from a combination of pathways and four main cell types: enterocyte, hepatocyte, macrophage, and erythroblast. The EGF/EGFR signaling pathway, heme production, STAT signaling, cAMP signaling, ferritin storage, and BMP-SMAD signaling are all involved in iron regulation. We conducted an exome-wide, gene-based recessive diplotype scan using putative functional variants to reveal additional genes underlying hemochromatosis susceptibility—an approach that can be widely applied to investigate complex disease susceptibility generated by compound heterozygosity and recessive single site effects using existing exome-wide association genotype and sequencing data. The scan identified *FGF6* as being significantly associated with hemochromatosis. *FGF6* belongs to the paracrine FGF-gene family and is largely expressed in skeletal muscle, which plays an important role in iron metabolism as it contains 10%-15% of iron stores. Functional experiments demonstrated that FGF-6 strongly impacted hepcidin expression to regulate iron homeostasis and decreases Fe2+ absorption in hepatocytes, while not impacting hepcidin-independent Fe3+ uptake. These results suggest FGF-6 mediates its effect on iron metabolism via hepcidin. The induction of hepcidin expression by FGF-6, presumably promotes ferroportin inhibition. We additionally found that three *FGF6* nonsynonymous variants increased intracellular Fe2+ concentrations and reduced hepcidin levels compared to wildtype *FGF6*, indicating loss-of-function. Interestingly, a genome-wide RNAi profiling study reported that knockdown of *FGF6* increased transferrin-mediated endocytosis.[36](#_ENREF_36) Rs12368351, approximately 8kb downstream of *FGF6* has been associated with phosphorus levels;37 and two SNPs, rs140668749 and rs10849061, within 20kb downstream of *FGF6*, are associated with migraine.38,39 Previous studies have indicated that iron plays a role in autoimmunity and a study examining pulmonary arterial hypertension in SSc noted iron deposition in lung elastin fibres and giant cells[37](#_ENREF_37), however, epidermal iron deposition in SSc has not been previously investigated. We observed that FGF-6 is involved with iron deposition in SSc and liver cancer. Together, these results demonstrate that fibroblast growth factor receptor (FGFRs) signaling through FGF-6 is a critically important mechanism in iron metabolism.

## Tables

**Table 1:** **Significant genes identified by recessive diplotype scanning.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CHR | GENE | P | OR# | SNPs | Case+ | Case- | Control+ | Control- |
| chr6 | *HFE* | 1.29×10−8 | 28.6 | 14 | 8 | 10 | 189 | 6707 |
| chr12 | *FGF6* | 1.99×10−6 | 22.8 | 10 | 6 | 12 | 153 | 6743 |
| chr21 | *KRTAP15-1* | 7.55×10−5 | 6.78 | 5 | 11 | 7 | 1271 | 5625 |
| chr20 | *XKR7* | 1.18×10−4 | 43.6 | 7 | 3 | 15 | 35 | 6861 |
| chr20 | *CABLES2* | 1.28×10−4 | 42.4 | 7 | 3 | 15 | 36 | 6860 |
| chr22 | *THOC5* | 1.38×10−4 | 6.24 | 9 | 13 | 5 | 1945 | 4951 |

The most significant six genes identified in the recessive diplotype test are displayed. P-values are from a two-tailed Fisher’s exact test. OR# shown as Haldane odds ratio. SNPs: number of genotyped variants per gene that were polymorphic in the samples studied. Case+: number of iron overload case individuals carrying recessive diplotypes with putative functional alleles. Case-: number of cases carrying at least one homolog at the gene without a putative functional allele. Control+: number of control individuals carrying recessive diplotypes with putative functional alleles. Control-: number of controls carrying at least one homolog at the gene without a putative functional allele.

## Figure legends

**Figure 1.** **Manhattan plot of the gene-based recessive diplotype association results.** The association *P*-value testing hemochromatosis association for each gene (y-axis) on different chromosomes is shown in alternating navy blue and green along the x-axis, with genes reaching experiment-wise significance for the gene-based analyses (experiment-wise  = 3.14 × 10−6) depicted in red.

**Figure 2.** **Comparative genomic analysis.** The comparative genomic analyses revealed that *FGF6* evolved synchronously with other iron metabolism genes. **(A)** Main iron metabolism genes were collected and alignment was conducted to make the comparative genomic analysis together with *FGF6*. The earliest gene appearance over time was inferred by comparing species and corresponding evolution and appearance time was labelled. **(B)** Protein-protein interaction network was estimated by String (version 10.0)[38](#_ENREF_38) using the highest confidence setting (confidence score>0.9). **(C)** Protein domain summarized by a previous FGF-6 functional study[39](#_ENREF_39). HBS: Heparin Binding Sites; FGFR-BR: FGFR-Binding Region.

**Figure 3.** **FGF-6 regulates hepcidin-dependent iron uptake.** Prussian blue staining was applied for the evaluation of iron deposition. Fe2+, Fe3+ as well as FGF-6 active proteins were added into the cell culture medians, respectively, with different recombination. Control group was treated with vehicle alone (boiled water). Iron staining was visualized by microscopy. Iron positive staining cells divided by total cell number was used to evaluate the iron deposition levels.

**Figure 4.** ***FGF6* loss-of-function nonsynonymous variants cause hepcidin downregulation and iron deposition. (A)** Prussian blue staining to HepG2 cells with different nonsynonymous mutations using Fe2+ and Fe3+ treatment. **(B)** quantitation of Iron absorption difference between *FGF6* plasmid transfection with wildtype, M1 (E127X), M2 (D174V), M3 (R188Q) mutation. Iron staining was visualized by microscopy. Iron positive staining cells divided by total cell number was used to evaluate the iron deposition levels. **(C)** Iron metabolism gene expression changes with FGF-6 active protein treatment in the cell culture media and *FGF6* mRNA transfection. **(D-G)** Iron metabolism gene expression changes after the transfection by *FGF6* mRNA into various cell types with wildtype and the identified variants R188Q, D174V and E172X. Cell lines: 786-O are kidney adenocarcinoma cells, A-498 are kidney carcinoma cells, HCT-8 are ileocecal colorectal adenocarcinoma cells, and HepG2 are liver hepatocellular carcinoma cells.

**Figure 5.** **Decreased *FGF6* expression gene in systemic sclerosis and hepatic cancer.** **(A)** FGF-6 protein level was evaluated by immunohistochemistry assay (IHC) in skin tissues from SSc patients. **(B)** Prussian blue staining was applied to evaluate the iron deposition in SSc skin tissues. **(C)** Immunohistochemistry assay of FGF-6 protein in liver cancer tissues. **(D)** Prussian blue staining was applied to evaluate the iron deposition in liver cancer tissues. FGF-6 expression level (Immunohistochemistry,IHC,staining) were counted by average optical density (AOD) of positive signal in each sample. Iron staining was visualized by microscopy. The ratio of iron-positive stain areas to the total area was used to evaluate the iron deposition levels by Image J software.

## Supplementary Figure Legends

**Supplementary Figure 1. Two-Site Power Calculations.** Power calculations for a two-site disease model comparing the Armitage trend test of disease association at each site to a log-likelihood ratio test explicity evaluating recessive diplotype effects. Baseline haplotype frequencies, case and control diploid sample sizes, and relative risk of disease-predisposing diplotypes parameters are shown. The of initial haplotype frequencies (A1B1, A1B2, A2B1, A2B2) are presented. Different combinations of haplotypes are generated by generating recombination between the two sites and the results are presented in a collapsed manner through a single linkage disequilibrium metric. Hardy-Weinberg equilibrium of haplotypes/diplotypes in the general population is assumed. R is the relative risk of disease for recessive diplotypes compared to the remaining diplotypes. ncs and nct are the number of cases and controls, respectively. The type 1 error rate, adjusted for an exome-wide scan, was set to 2.5E-06 for all calculations.

**Supplementary Figure 2.** **Quantile-Quantile Plot.** Q-Q plot for the exome-wide, gene-based recessive diplotype scanning in hemochromatosis is shown. Numerous genes had no recessive diplotypes with putative functional alleles and therefore yielded P-values of 1. The two data points exceeding the confidence interval represented *HFE* and *FGF6*.

**Supplementary Figure 3.** **Time course for response to FGF-6 protein treatment.** HepG2 cells were treated with FGF-6 protein for different times and iron metabolism gene expression determined. LL, L and H indicate low, mediate and high level of FGF6 protein level.

**Supplementary Figure 4**. **Three nonynonymous *FGF6* mutations and PCR primer construction systems.**

The E172X (M1), D174V (M2), and R188Q (M3) plasmid constructs are shown along with the forward and reverse primers.

**Supplementary Figure 5**. **FGF-6 expression in normal hepatocytes, non-metastatic liver cancer cells and metastatic liver cancer cells.** **(A)** FGF-6 expression in normal cells and metastatic cancer cells. However, metastatic liver cancer cells (red arrow in **A**) showed higher FGF-6 expression compared with normal hepatocytes and non-metastatic liver cancer cells **(B)**.

## Supplementary Table Legends

**Supplementary Table 1.** **Nonsynomymous single nucleotide variants (SNV) in *FGF6* exon regions.** The table displays each of the ten nonsynonymous variants interrogated in the study with genomic position, corresponding reference and alternative alleles, SNP type, and corresponding amino acid changes. All SNVs were on the exome beadchip array applied to the sample set.

**Supplementary Table 2.** **RT-PCR primers for six iron metabolism genes.** Information for the RT-PCR assays are shown with TM values and amplicon size for the six iron metabolism genes studied.

## Online Methods

#### Marshfield PMRP Hemochromatosis Dataset

The homogenous population in rural Central Wisconsin is the source population for the Personalized Medicine Research Project (PMRP), a biobank linked to electronic health records (EHRs) housed by the Marshfield Clinic Research Institute.40 Samples from over 20,000 individuals comprise the PMRP. The Central Wisconsin population is largely stationary and primarily derived from Bavarian migrants in the late 1800s. The population carries high utility for disease gene mapping through reduction in confounding by population stratification and lower expected levels of allelic and locus heterogeneity. In addition, environmental exposures are thought to be relatively uniform across this population. For these reasons, the PMRP has been effectively used in numerous human genetics studies.[40-43](#_ENREF_40) PMRP DNA samples were collected and stored approximately 14 years ago and all individuals have longitudinal EHR information housed at the Marshfield Clinic, averaging in excess of 30 years. The EHR is composed of ICD-9 diagnostic codes, laboratory test results, clinical procedure data, prescription information and physician notes. Hemochromatosis cases and controls were selected from the PMRP population. PMRP individuals were selected as hemochromatosis cases on the basis of percent transferrin saturation laboratory values (the ratio of serum iron to transferrin iron-binding capacity) exceeding 48% and having two or more instances of ICD-9 codes indicating the diagnosis of hemochromatosis: 275.0 (iron metabolism disorder, excluding anemia), 275.01 (hereditary hemochromatosis), 275.03 (unspecified hemochromatosis), and/or 275.09 (other iron metabolism disorders). Using a PCA analysis of the exome genotyping data, individuals considered genetic background outliers were excluded from the study. Of the approximately 10,000 individuals previously subjected to the exome genotyping array, 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.

#### Genotyping Platform

Of the full PMRP cohort, approximately 10,000 DNAs were interrogated by high density genotyping on the Illumina HumanCoreExome beadchip. The exome array has ancestry informative markers, a panel of identity-by-descent SNPs, coverage of markers found to be genome-wide significant in GWAS studies, and excellent coverage of exonic variants. Over 500,000 variants are interrogated on this genotyping beadchip. The version of the beadchip was designed and used in the AMD consortium.44 Additional recent studies have used data generated from this genotyping platform to discover susceptibility genes for common diseases.46

#### Determination of Putative Functional Variants

Putative functional variants included in the analyses satisfied the quality control criteria as described previously.44 In addition, markers in the analyses were either GWAS-significant as of June 2015 and/or annotated as missense, nonsense, 3’UTR, 5’UTR or occurring within a splice site region.

#### Haplotype Phasing

In general, gametic phasing is necessary to directly determine compound heterozygous individuals at a particular gene. Beagle is applied for inferring phased haplotypes from unphased genotype data using population-based samples, including the localized haplotype-cluster model algorithm.[44](#_ENREF_44) Each gene in the exome was phased separately using this method within Beagle. Phasing was performed using all cases and controls in combination. 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.43 Although phasing accuracy is generally compromised with rare variants, it should be noted that randomly assigning alleles to a homolog in individuals carrying multiple putative functional alleles at a gene yields a probability that the individual has a recessive diplotype configuration of , where is the number of putative functional variants. Hence, an individual carrying four, six or eight putative functional variants at a gene would have a probability of carrying a recessive diplotype of 0.875, 0.969 and 0.992, respectively, given random phasing.

#### Statistical Tests of Recessive Diplotypes

At each gene, individuals were classified as having a recessive diplotype configuration if they carried at least one putative functional allele on each homolog (). Individuals carrying at least one homolog free from putative functional alleles were deemed as having a wildtype diplotype (). The total number of case and control individuals carrying a recessive diplotype was denoted by and , respectively. Similarly, the total number of case and control individuals carrying a wildtype diplotype was denoted by and , respectively. Following the determination of these counts, a Fisher’s exact test was applied to the contingency table. Individuals carrying one or more homozygous genotype(s) at a single site for a putative functional allele were included in the category. Genes without any high quality, putative functional alleles across all samples were removed from the analyses. Across all genes with analyzable data, a conservative experiment-wise correction was used to correct for multiplicity using 15,900 gene-based tests.

#### Power Calculations

To explore the efficacy of the approach proposed here, we performed analytic power calculations under the alternative model of compound heterozygosity/recessive inheritance of disease at two sites, each segregating two alleles. By doing so, we sought to compare the power of a standard GWAS analysis (Armitage trend test) to a log-likelihood ratio test. **Supplemental Figure 1** shows the power of each of these tests across different sets of penetrances and haplotype frequencies. To consolidate the different sets of haplotype frequencies, the results are plotted as a function of linkage disequilibrium between the two sites. The power of the test for recessive diplotypes exceeded the power for the Armitage trend test across virtually all of the parameter space. Additional work in this area was recently performed by Sanjak and colleagues showing similar results.[45](#_ENREF_45)

#### Comparative genomic analysis and Protein-protein interaction inference

Amino acid sequencing of the core iron metabolism genes were collected including *TFRC, FTH1, IREB1, SKP1, SKP1, ACO1, TFR2, TF, HMOX1, ACO2, HAMP, FGF6, and FGFR1*. The alignments were derived from NCBI BLASTn database. Phylogeny for different genes were compared to show the earliest evolutionary time point and then occurrence for each gene were mapped to the phylogenetic tree.[46](#_ENREF_46) Protein-protein interaction network inference was conducted to FGF-6 and main iron metabolism genes. The finial network was tuned after removing non-necessary nodes between FGF-6 and key iron molecules including HFE and SLC40A1.

#### Cell culture, Reagents, and Protein treatment

Colon cancer cell lines HCT-8, Kidney cancer cell lines A498, 786-O and Liver cancer cell line HepG2 were cultured in DMEM medium supplemented with 10% FCS at 37°C in a 5% CO2 humidified incubator. Recombinant human FGF-6 protein (Active) (ab219122) was purchased from Abcam. FeSO4·7H2O, Fe2(SO4)3·XH2O and a mixture of FeSO4·7H2O and Fe2(SO4)3·XH2O was dissolved in 50 ml of water already boiled to eliminate oxygen. Plasmid with raw *FGF6* sequence was purchased from PPL-Shanghai Co., Ltd (Shanghai, China) while 3 different mutations (E127X, D174V,R188Q) were synthesized with overlapping-PCR.

#### Prussian blue 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 microscopy. Iron positive staining cells divided by total cell number was used to evaluate the iron deposition levels.

#### RT-PCR and quantitative RT-PCR analysis

Total RNA was extracted from the cells using Trizol (Invitrogen, Carlsbad, CA, USA). One microgram of total RNA was subjected to cDNA synthesis using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to the manufacturer’s instructions. The specific primers for each gene were designed using Primer 5 and synthesized by Generay Biotech Co., Ltd. (Shanghai, China). The RT-PCR amplification was conducted using a SYBR Green I PCR Kit (TaKaRa, Shanghai, Japan) according to manufacturer’s instructions. The reaction was carried out on a ABI Prism 7900 Detector System (Applied Biosystems). RT-PCR conditions were 95°C for 3 min, followed by 40 cycles of 95°C for 15 s, 60°C for 40 s, and the conditions for obtaining the dissociation curve were 95°C for 15 s, 60°C for 15 s, 95°C for 15 s. The data obtained from the assays were analyzed with SDS 2.3 software (Applied Biosystems). For each sample, the relative gene expression was calculated using a relative ratio to GAPDH. Related RT-PCR primers can be found in the **Supplementary Table 2**.

#### Immunohistochemical staining of FGF-6

The primary antibody used was anti-FGF6 (1:200, D162668 BBI, Shanghai). Liver and skin tissues from four liver cancer patients and six SSc patients, respectively, and normal controls were formalin-fixed and paraffin-embedded. Sections were deparaffinized and incubated with 5% bovine serum albumin for 60 min. Cells positive for FGF6 were detected by incubation with the primary antibody for 2 h at room temperature followed by incubation with 3% hydrogen peroxide for 10 min. Rabbit anti-rabbit lgG labeled with horseradish peroxidase were used as secondary antibodies. The expression of FGF6 was visualized with 3,3-diaminobenzidinetetrahydrochloride (DAB-4HCl). The expression of FGF-6 in SSc and tumor tissues were quantitated by the average optical density (AOD) of positive signal in each sample using the software imageJ (Windows and Java-1.8.0, NIH).

## Acknowledgements

on April 15, 2014, the study was reviewed and approved by the Marshfield Clinic Research Institute Institutional Review Board. FWA00000873, IRB00000673, Title: *Two Allele Loss of Function Genotype Array Study*. The project described was supported by the Clinical and Translational Science Award (CTSA) program, previously through the National Center for Research Resources (NCRR) grant 1UL1RR025011 and the National Center for Advancing Translational Sciences (NCATS) grant 9U54TR000021, and now by the NCATS grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additional funding was provided by the Marshfield Clinic Research Institute grant SCH10218 and generous donors to the Marshfield Clinic.

## Author contributions

SG performed analyses, interpreted results, designed the functional studies, and aided in drafting the manuscript. SJ conducted molecular and cell biology experiments. MM aided in the analyses and reviewed the manuscript. MW, YM and WW provided clinical and biochemistry advice and aided in drafting the manuscript. MM aided with the experimental design and analyses. ZY performed initial genetic and statistical analyses, performed data management and reviewed the manscript. BO implemented and refined the phenotyping algorithms. TK and JJ aided in the regulatory paperwork and reviewed the manuscript. RS performed data management tasks. JJM provided clinical advice and reviewed the manuscript. JKM supervised the management of biological samples for genotyping and reviewed the manuscript. LJ reviewed the manuscript and provided general scientific advice. JAS provided molecular and cellular biology advice, clinical advice, reviewed and edited the manuscript. JW supervised the functional studies, reviewed the manuscript and provided biological advice. SJS designed the experiment, supervised the genetic analyses, developed phenotyping algorithms, developed analysis methods and power calculations, interpreted results and aided in drafting and editing the manuscript.

## Analysis Code:

We uploaded the Code to github: <https://github.com/Shicheng-Guo/marshfield/blob/master/2ALOF/readme.md>

## References

1. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).

2. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).

3. Zou, J. *et al.* Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat Commun* **7**, 13293 (2016).

4. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).

5. Andreoletti, G. *et al.* Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway. *Sci Rep* **7**, 46454 (2017).

6. De Braekeleer, M., Allard, C., Leblanc, J.P., Simard, F. & Aubin, G. Genotype-phenotype correlation in cystic fibrosis patients compound heterozygous for the A455E mutation. *Hum Genet* **101**, 208-11 (1997).

7. Prietsch, V. *et al.* Mevalonate kinase deficiency: enlarging the clinical and biochemical spectrum. *Pediatrics* **111**, 258-61 (2003).

8. Thein, S.L. Genetic modifiers of beta-thalassemia. *Haematologica* **90**, 649-60 (2005).

9. Bauer, P. *et al.* NPC1: Complete genomic sequence, mutation analysis, and characterization of haplotypes. *Hum Mutat* **19**, 30-8 (2002).

10. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* **19**, 571-7 (2016).

11. Adam, R. *et al.* Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *Am J Hum Genet* **99**, 337-51 (2016).

12. Hague, S. *et al.* Early-onset Parkinson's disease caused by a compound heterozygous DJ-1 mutation. *Ann Neurol* **54**, 271-4 (2003).

13. Onoufriadis, A. *et al.* Mutations in IL36RN/IL1F5 are associated with the severe episodic inflammatory skin disease known as generalized pustular psoriasis. *Am J Hum Genet* **89**, 432-7 (2011).

14. Dewey, F.E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**(2016).

15. Adams, P.C. & Barton, J.C. Haemochromatosis. *Lancet* **370**, 1855-60 (2007).

16. Andrews, N.C. & Schmidt, P.J. Iron homeostasis. *Annu Rev Physiol* **69**, 69-85 (2007).

17. Saddi, R. & Feingold, J. Idiopathic haemochromatosis: an autosomal recessive disease. *Clin Genet* **5**, 234-41 (1974).

18. Borecki, I.B., Rao, D.C., Yaouanq, J. & Lalouel, J.M. Segregation of genetic hemochromatosis indexed by latent capacity of transferrin. *Am J Hum Genet* **45**, 465-70 (1989).

19. Simon, M., Alexandre, J.L., Bourel, M., Le Marec, B. & Scordia, C. Heredity of idiopathic haemochromatosis: a study of 106 families. *Clin Genet* **11**, 327-41 (1977).

20. Cartwright, G.E. *et al.* Inheritance of hemochromatosis: linkage to HLA. *Trans Assoc Am Physicians* **91**, 273-81 (1978).

21. Edwards, C.Q., Griffen, L.M., Dadone, M.M., Skolnick, M.H. & Kushner, J.P. Mapping the locus for hereditary hemochromatosis: localization between HLA-B and HLA-A. *Am J Hum Genet* **38**, 805-11 (1986).

22. Jazwinska, E.C., Lee, S.C., Webb, S.I., Halliday, J.W. & Powell, L.W. Localization of the hemochromatosis gene close to D6S105. *Am J Hum Genet* **53**, 347-52 (1993).

23. Feder, J.N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* **13**, 399-408 (1996).

24. Jazwinska, E.C. *et al.* Haemochromatosis and HLA-H. *Nat Genet* **14**, 249-51 (1996).

25. Griffiths, W. & Cox, T. Haemochromatosis: novel gene discovery and the molecular pathophysiology of iron metabolism. *Hum Mol Genet* **9**, 2377-82 (2000).

26. Allen, K.J. *et al.* Iron-overload-related disease in HFE hereditary hemochromatosis. *N Engl J Med* **358**, 221-30 (2008).

27. Milet, J. *et al.* Common variants in the BMP2, BMP4, and HJV genes of the hepcidin regulation pathway modulate HFE hemochromatosis penetrance. *Am J Hum Genet* **81**, 799-807 (2007).

28. Benyamin, B. *et al.* Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. *Nat Commun* **5**, 4926 (2014).

29. de Tayrac, M. *et al.* Genome-wide association study identifies TF as a significant modifier gene of iron metabolism in HFE hemochromatosis. *J Hepatol* **62**, 664-72 (2015).

30. Recalcati, S., Locati, M., Gammella, E., Invernizzi, P. & Cairo, G. Iron levels in polarized macrophages: regulation of immunity and autoimmunity. *Autoimmun Rev* **11**, 883-9 (2012).

31. Bowlus, C.L. The role of iron in T cell development and autoimmunity. *Autoimmun Rev* **2**, 73-8 (2003).

32. Dixon, S.J. *et al.* Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* **149**, 1060-72 (2012).

33. Torti, S.V. & Torti, F.M. Iron and cancer: more ore to be mined. *Nat Rev Cancer* **13**, 342-55 (2013).

34. Derynck, R., Akhurst, R.J. & Balmain, A. TGF-beta signaling in tumor suppression and cancer progression. *Nat Genet* **29**, 117-29 (2001).

35. Wakefield, L.M. & Roberts, A.B. TGF-beta signaling: positive and negative effects on tumorigenesis. *Curr Opin Genet Dev* **12**, 22-9 (2002).

36. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464**, 243-9 (2010).

37. Overbeek, M.J. *et al.* Pulmonary arterial hypertension in limited cutaneous systemic sclerosis: a distinctive vasculopathy. *Eur Respir J* **34**, 371-9 (2009).

38. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-52 (2015).

39. Li, Y. *et al.* Heparin binding preference and structures in the fibroblast growth factor family parallel their evolutionary diversification. *Open Biol* **6**(2016).

40. Carter, T.C. *et al.* Validation of a metabolite panel for early diagnosis of type 2 diabetes. *Metabolism* **65**, 1399-408 (2016).

41. Ye, Z. *et al.* Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to Staphylococcus aureus infections. *Front Genet* **5**, 125 (2014).

42. Ritchie, M.D. *et al.* Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* **127**, 1377-85 (2013).

43. Fritsche, L.G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* **48**, 134-43 (2016).

44. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).

45. Sanjak, J.S., Long, A.D. & Thornton, K.R. A Model of Compound Heterozygous, Loss-of-Function Alleles Is Broadly Consistent with Observations from Complex-Disease GWAS Datasets. *PLoS Genet* **13**, e1006573 (2017).

46. dos Reis, M. *et al.* Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr Biol* **25**, 2939-50 (2015).