scientific reports

OPEN

The diagnostic yield of exome sequencing in liver diseases from a curated gene panel

Xiao-Fei Kong[1,2,3,6]✉, Kelsie Bogyo[2], Sheena Kapoor[2], Patrick R. Shea[2,3], Emily E. Groopman[2], Amanda Thomas-Wilson[4,7], Enrico Cocchi[2], Hila Milo Rasouly[2], Beishi Zheng[1], Siming Sun[1], Junying Zhang[2], Mercedes Martinez[5], Jennifer M. Vittorio[5,8], Lorna M. Dove[1,5], Maddalena Marasa[2], Timothy C. Wang[1], Elizabeth C. Verna[1,5], Howard J. Worman[1,4], Ali G. Gharavi[2,3,9], David B. Goldstein[3,9] & Julia Wattacheril[1,5,9]✉

Exome sequencing (ES) has been used in a variety of clinical settings but there are limited data on its utility for diagnosis and/or prediction of monogenic liver diseases. We developed a curated list of 502 genes for monogenic disorders associated with liver phenotypes and analyzed ES data for these genes in 758 patients with chronic liver diseases (CLD). For comparison, we examined ES data in 7856 self-declared healthy controls (HC), and 2187 patients with chronic kidney disease (CKD). Candidate pathogenic (P) or likely pathogenic (LP) variants were initially identified in 19.9% of participants, most of which were attributable to previously reported pathogenic variants with implausibly high allele frequencies. After variant annotation and filtering based on population minor allele frequency (MAF ≤ 10^{-4} for dominant disorders and MAF ≤ 10^{-3} for recessive disorders), we detected a significant enrichment of P/LP variants in the CLD cohort compared to the HC cohort (X^2 test OR 5.00, 95% CI 3.06–8.18, *p* value = 4.5e−12). A second-level manual annotation was necessary to capture true pathogenic variants that were removed by stringent allele frequency and quality filters. After these sequential steps, the diagnostic rate of monogenic disorders was 5.7% in the CLD cohort, attributable to P/LP variants in 25 genes. We also identified concordant liver disease phenotypes for 15/22 kidney disease patients with P/LP variants in liver genes, mostly associated with cystic liver disease phenotypes. Sequencing results had many implications for clinical management, including familial testing for early diagnosis and management, preventative screening for associated comorbidities, and in some cases for therapy. Exome sequencing provided a 5.7% diagnostic rate in CLD patients and required multiple rounds of review to reduce both false positive and false negative findings. The identification of concordant phenotypes in many patients with P/LP variants and no known liver disease also indicates a potential for predictive testing for selected monogenic liver disorders.

Abbreviations

ES	Exome sequencing
CLD	Chronic liver disease
CKD	Chronic kidney disease
HGMD	Human Gene Mutation Database
PTV	Protein truncating variants

[1]Division of Digestive and Liver Diseases, Department of Medicine, Columbia University Irving Medical Center, Hammer Health Sciences Building Rm 402, 701 W 168th St, New York, NY 10032, USA. [2]Center for Precision Medicine and Genomics, Department of Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA. [3]Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA. [4]Department of Pathology and Cell Biology, Columbia University Irving Medical Center, New York, NY 10032, USA. [5]Center for Liver Disease and Transplantation, Columbia University Irving Medical Center, 622 West 168th Street, PH 14-105D, New York, NY 10032, USA. [6]Present address: Department of Medicine, McDermott Center for Human Growth and Development, UT Southwestern Medical Center, Dallas, TX 75390-9151, USA. [7]Present address: Molecular Diagnostics, New York Genome Center, New York, NY, USA. [8]Present address: NYU Transplant Institute, NYU Langone Health, New York, NY, USA. [9]These authors contributed equally: Ali G. Gharavi, David B. Goldstein and Julia Wattacheril. ✉email: xiao-fei.kong@utsouthwestern.edu; jjw2151@cumc.columbia.edu

MAF	Minor allele frequency
NGS	Next-generation sequencing
GWAS	Genome-wide association studies
ACMG-AMP	American College of Medical Genetics and Genomics and the Association for Molecular Pathology
OMIM	Online Mendelian Inheritance in Man
HPO	Human Phenotype Ontology
gnomAD	Genome Aggregation Database
ExAC	Exome Aggregation Consortium
AD	Autosomal dominant
AR	Autosomal recessive
XLD	X-linked dominant
XLR	X-linked recessive disorders
DM	Disease-causing mutation
P/LP	Pathogenic/likely pathogenic
CADD	Combined Annotation Dependent Depletion
QC	Quality control

Liver disease accounts for approximately 2 million deaths per year worldwide. In the United States, the mortality rate for chronic liver diseases (CLD) increased 31% from 2000 to 2015, making it the fifth leading cause of death in 2017 for persons aged 45–64 years[1]. The history of liver genetic diseases dates back to 1865–1890 when Triouseau and von Recklinghausen described hemochromatosis[2]. The cloning, mapping, and functional characterization of homeostatic iron regulator (*HFE*) gene in the 1990s paved the way for molecular diagnosis of hemochromatosis[3]. The advent of next-generation sequencing (NGS) approaches have led to the discovery of genetic disorders causing liver disease phenotypes such as fibrolamellar hepatocellular carcinoma[4], recurrent acute liver failure[5–7], or idiopathic non-cirrhotic portal hypertension[8,9]. These findings demonstrate the power of NGS for identifying novel genetic forms of liver diseases.

NGS has been successfully deployed in clinical care to diagnose monogenic forms of neurologic, developmental, cardiac or renal disorders[10,11]. While genetic testing of single genes or small gene panels has been used for some suspected hereditary liver diseases[12–16], NGS approaches have not been widely adopted into the routine evaluation of liver disease. As sequencing costs decline and clinical utility is demonstrated, a standardized genetic diagnostic pipeline for liver disease could benefit patients and clinicians, enabling efficient clinical diagnoses and early recognition of rare genetic disorders that may manifest as a common liver phenotype and may not be recognized based on their clinical workup[17]. In this paper, we outline an analytic approach and conduct a clinical sequence interpretation for ES data from 10,801 individuals (Supplementary Table 1), including 758 patients with CLD as encountered at various stages of their diagnostic workflow. Here we present the diagnostic utility of ES for liver diseases, highlight special considerations and elaborate on the potential for misclassification in the genetic workup for liver diseases.

Results

Characterization of 502 liver genes with Mendelian hepatobiliary disorders

In a comprehensive search for Mendelian genetic disorders with any liver abnormalities prompting clinical referral to a hepatologist, we manually curated a total of 959 genes. Of these, 502 had a confirmed abnormal and broad liver disease phenotype, with 193 genetic disorders having primarily liver disease. For example, *ABCB11* or *ATP8B1* causing progressive familial intrahepatic cholestasis; other genes might lead to liver abnormalities that are presenting clinically as a secondary cause. For instance, patients with Fanconi anemia may present with hepatocellular carcinoma[18], individuals with inborn errors of immunity may have acute or chronic liver infection[19] (Supplementary Fig. 1). We then annotated inheritance modes and detailed clinical phenotypes related to these 502 genes. In total, 75% of genes were associated with a recessive mode of inheritance (363 AR and 15 XLR). Sixty-two autosomal genes could result in both dominant and recessive disorders and 62 other genes associated with exclusively dominant disorders (61 AD and 1 XLD) (Supplementary Table 2). The most common clinical presentation of Mendelian hepatobiliary disorders was hepatomegaly, manifesting in 236/502 disorders (47%) Other common clinical manifestations included metabolic disease (25%), liver fibrosis or cirrhosis (25%), elevated hepatic transaminase level (20%), and cholestasis (19%) (Fig. 1A). Most of the genes (298/502, 59%) were associated with a developmental or congenital disorder with liver manifestations (Supplementary Table 2). The 62 genes exclusively associated with dominant inheritance showed significantly higher pLI (Fig. 1B) and missense Z scores (Fig. 1C) compared to the 378 genes associated with recessive diseases. For 62 genes associated with both dominant and recessive inheritance, a total of 16 genes has pLI score above 0.8 and nine of sixteen were involved in the immune system. Two genes, *STAT1* and *INSR*, have a missense Z score above three (Supplementary Information 1). In conclusion, we curate a total of 502 genes that can be related to liver phenotypes in a Mendelian inheritance and evaluate its value in ES data analysis.

Assessment of the frequency of candidate pathogenic/likely pathogenic variants

To investigate the prevalence of candidate pathogenic variants in the liver genes, we analyzed ES data from 10,801 individuals, agnostic to the clinical phenotype. 758 patients were diagnosed with CLD. In additional, two control cohorts were used to evaluate the gene-list based ES analysis, including 7856 self-identified healthy individuals and 2,187 patients from CUIMC with chronic kidney disease (CKD) (Supplementary Table 1 and Supplementary Table 3)[20,21]. Based on an automated filtering (DP > 9, VQSR filter = PASS, Qual > 49, QD > = 2,

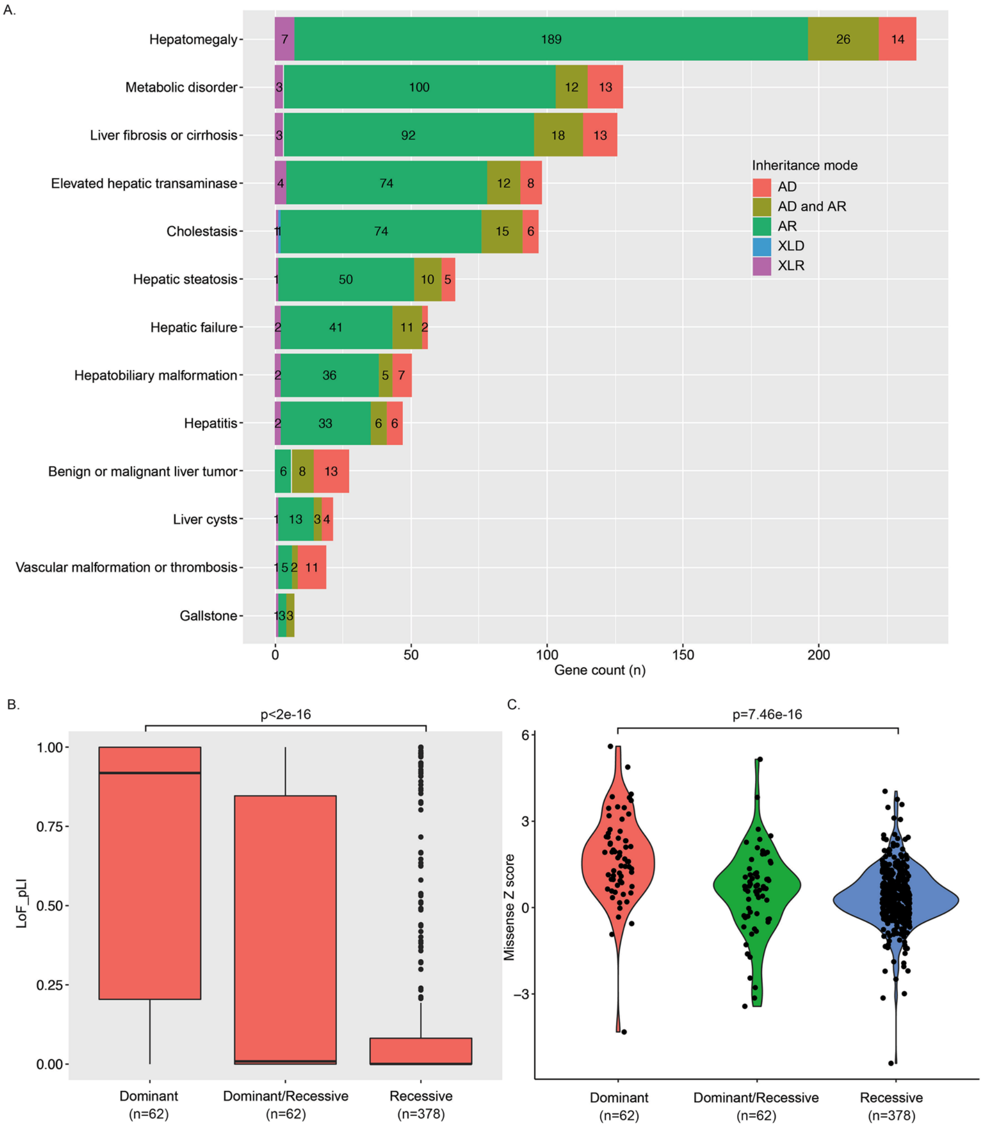

Figure 1. A summary of liver phenotypes in Mendelian genetic disorders. (**A**) Inheritance mode, annotated clinical liver phenotypes, and biological effects of 502 genes related to Mendelian disorders. The liver phenotypes and inheritance were curated based on OMIM, ClinGen, and a literature search. AD: autosomal dominant disorder; AR: autosomal recessive disorder; XLD: X-linked disorder; XLR: X-linked recessive disorder; AD and AR: Genes with both autosomal dominant and autosomal recessive inheritance were reported. The right lower box showed the numbers of genes with corresponding biological effects and inheritance mode. (**B**) Box plot of pLI scores of 502 genes in three groups based on inheritance mode. The dark line inside the box represents the median of pLI score. The top of box is 75% and bottom of box is 25%. The endpoints of the lines are at a distance of 1.5*IQR, where inter quartile range is the distance between 25 and 75th percentiles. The points outside the whiskers are marked as dots and are considered as extreme points. (**C**) Violin plot of missense Z scores of 502 genes in three different groups based on inheritance mode. *P* values in B and C for differences between dominant and recessive genes were determined using ANOVA.

GQ > = 20, MQ > = 40, Percentage of alt read > 0.25, MAF < 0.01)[20], we initially identified an equal distribution of candidate pathogenic variants, either "DM" in HGMD, or "Pathogenic" in ClinVar, across the three cohorts: 1567 (20.2%) in healthy controls, 416 (19.0%) in the CKD cohort, and 159 (21%) in the CLD cohort (Fig. 2A,B, Supplementary Table 4). This implausibly high frequency of variants for monogenic liver disorders suggested variant misclassification. Consistent with this conjecture, an analysis of CADD score and the maximal MAF from the ExAC and gnomAD indicated that many of these variants had implausibly high allele frequencies to be disease causing and had been erroneously reported as pathogenic prior to the availability public variant databases[22,23] (Fig. 2C, Supplementary Information 2). We next used the maximal MAF, MAF $\le 10^{-4}$ for dominant disorders and MAF $\le 10^{-3}$ for recessive disorders, to filter variants, followed by manual review of 403 variants (Fig. 2A,B, Supplementary Table 5)[20,23,24]. This resulted in 112 variants being classified as either P/LP based on

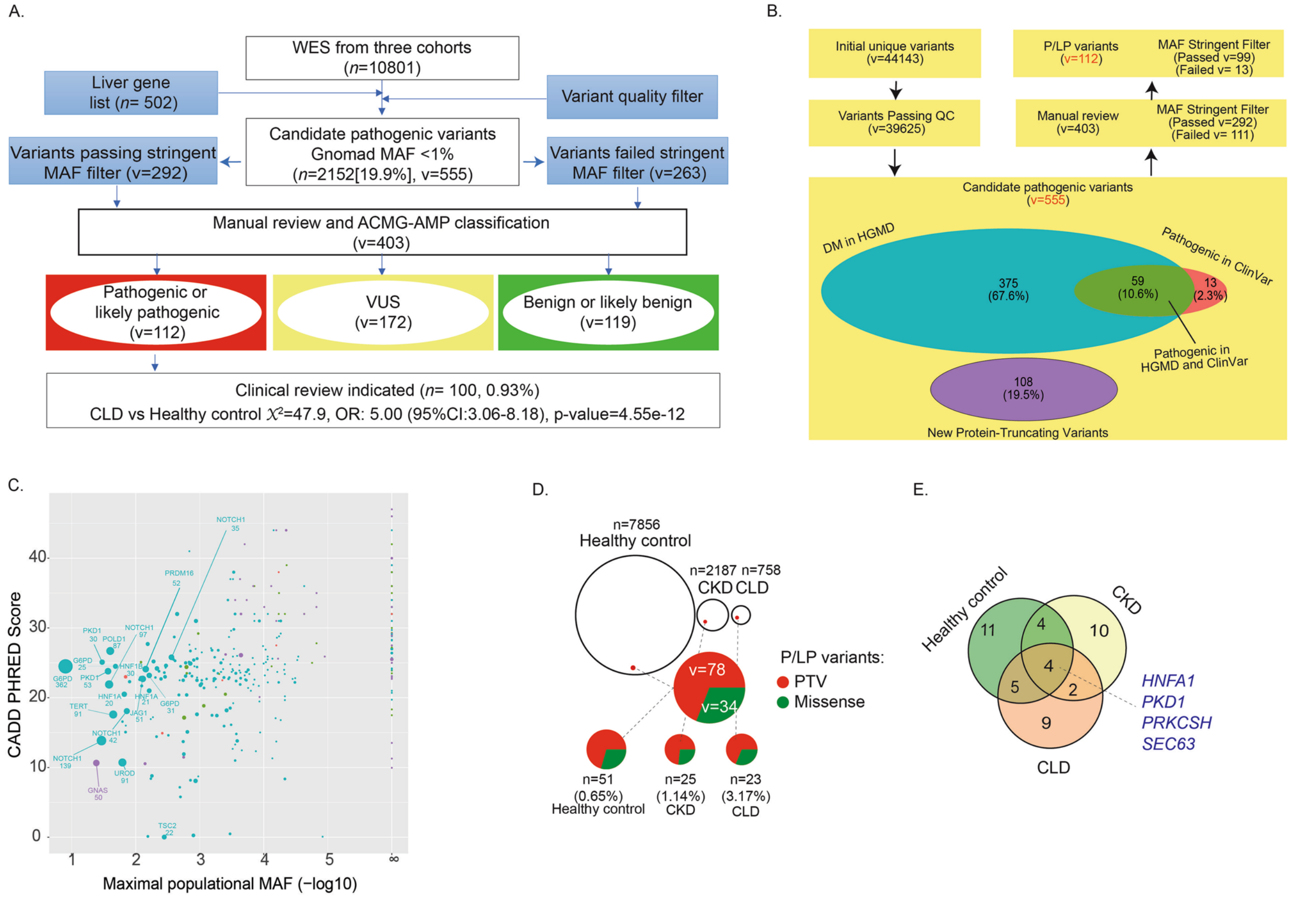

Figure 2. A search for candidate variants and ACMG-AMP classification revealed an enrichment of P/LP variants in the CLD cohort. (**A**) Approach used to identify pathogenic/likely pathogenic variants in the CLD cohort. We started with a search for all the candidate pathogenic variants with global AF less than 1% in gnomAD for 10,801 WES samples and ended up with an implausibly high frequency of monogenic disorders. We then applied a stringent filter based on maximal populational MAF and manually annotated a total of 403 variants based on ACMG-AMP guidelines and concluded that a total of 112 variants are pathogenic/likely pathogenic, and 1% of individuals might benefit from a further clinical evaluation. (**B**) A diagram for variants filtering and candidate pathogenic variants search for monogenic liver disease genes. Variants were classified based on the following: DM in HGMD but not pathogenic in ClinVar (cyan); Pathogenic in ClinVar but not DM in HGMD (Orange); Pathogenic in ClinVar and DM in HGMD (green); and new protein-truncating variants not reported in HGMD or ClinVar (purple). (**C**) Variants with high populational MAF in dominant disorders with liver phenotypes: X-axis is CADD Phred score of each variant; Y-axis is the -log10 of the highest MAF, which was extracted from the following subpopulations: African/African American (AFR), Latino (AMR), Ashkenazi Jewish (ASJ), Finnish (FIN), Non-Finnish European (NFE), East Asian (EAS), South Asian (SAS) and Other (OTH) from ExAC and gnomAD data. Circle size indicates the total number of individuals carrying the variant. If 20 or more individuals were found to be carriers, the gene name and count are given. (**D**) Schematic presentation of individuals in each cohort with pathogenic/likely pathogenic variants, the majority of which are PTVs. (**E**) A Venn diagram shows a total of 45 genes found in at least one affected individual from three cohorts. Five genetic disorders were found in all three cohorts.

ACMG-AMP classification (including 78 PTVs, Fig. 2D), detected across 45 genes in a total of 100 individuals (0.93% of three cohorts). Subsequent to this filtering and manual annotation process, the prevalence of these P/LP variants significantly differed between healthy controls (51/7856, 0.65%), patients with CKD (25/ 2187, 1.14%), and patients with CLD (24/758, 3.17%) (X^2 test OR: 5.00, 95%CI 3.06–8.18, p value = 4.55e–12, Fig. 2A,E). In summary, a search for rare variants in 502 genes associated with liver phenotypes lead to a significant enrichment of P/LP variants in the CLD cohort.

Second-level annotation of the CLD cohort identifies additional pathogenic variants

To maximize the identification of diagnostic variants in the CLD cohort, we performed a second-level manual assessment, using the more relaxed sequence quality thresholds which we had previously deployed to optimize diagnostic yield in other cohorts[21,25]. This second-level analysis led to the identification of 16 additional diagnostic variants that explained the liver phenotypes in 14 additional patients (13 genes, Fig. 3A). All 16 variants were missed because of the high stringency sequence quality thresholds and were all confirmed by Sanger sequencing.

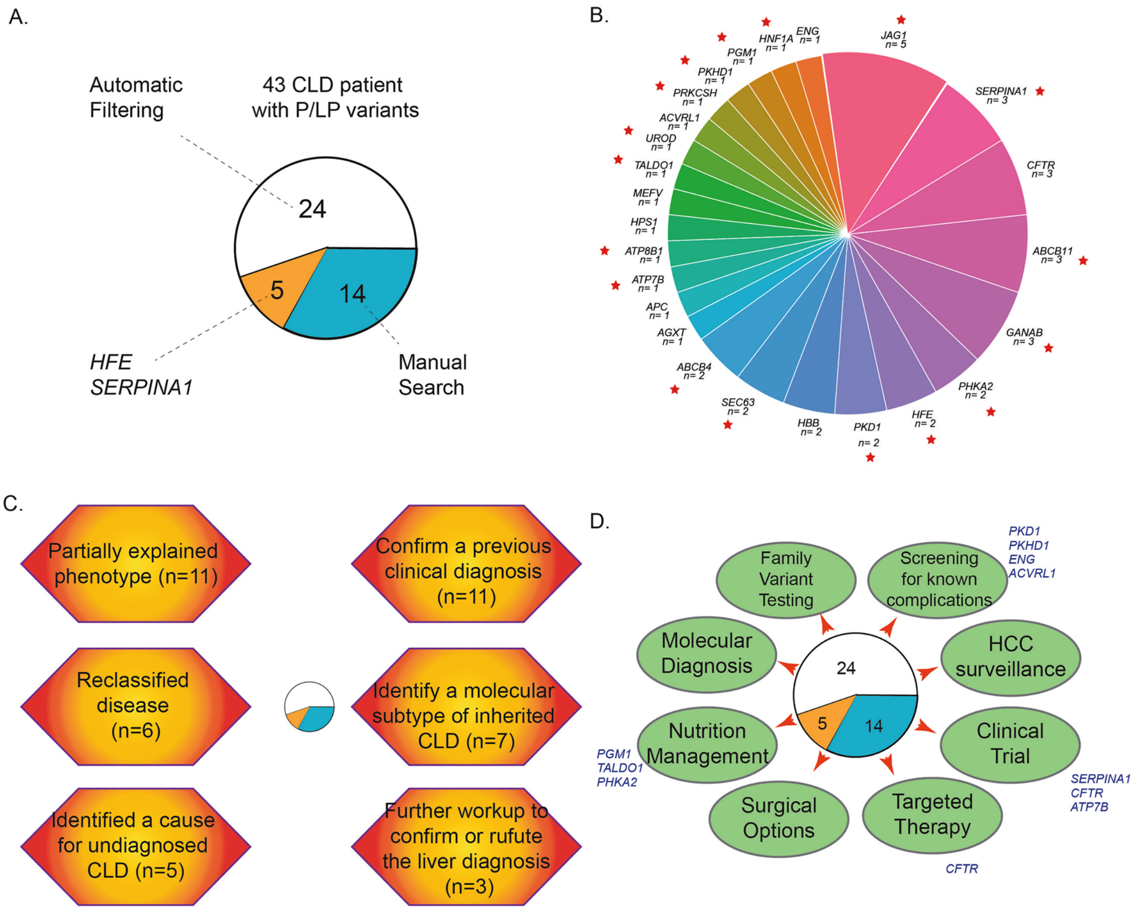

Figure 3. Genetic diagnoses and clinical implications of ES findings in the liver disease cohort. (**A**) A total of 43 CLD patients with P/LP variants from three searching approaches; (**B**) A total of 25 genetic disorders were found in the CLD cohort. Red star indicated the genetic disorders causing primarily liver diseases; (**C**) An investigation of clinical phenotypes and genetic diagnosis in CLD patients with P/LP variants; (**D**) clinical implications of the genetic findings.

In addition, we evaluated four well-known pathogenic variants or risk alleles for liver disease that have a MAF above 1%: *HFE* C282Y and H63D, *SERPINA1*E264V (Pi*S) and E342K (Pi*Z). We found two patients with P/LP variants in *HFE* (one with a homozygous *HFE* C282Y variant, and one with an H63D/c.340 + 1G > A genotype, Table 1). Both had high serum iron transferrin saturation and ferritin levels, and clinical presentations consistent with hereditary hemochromatosis. For *SERPINA1*, three patients in the CLD cohort had a homozygous Pi*Z genotype, and all of them had a clinical diagnosis of alpha-1 anti-trypsin deficiency (Table 1). Altogether, this second level analysis increased the diagnostic yield in the liver cohort to 43/758 cases (5.7%, Fig. 3A).

Genetic diagnoses and their clinical implications

Overall, we identified a total of 25 genetic disorders in the liver disease cohort, with Alagille syndrome, alpha-1 anti-trypsin deficiency, cystic fibrosis, and progressive familial intrahepatic cholestasis-2 detected in at least three patients each (Fig. 3B). There were no differences observed in sex, race, or ethnicity between the patients with or without a genetic diagnosis in the liver disease cohort. From a univariate analysis, younger age and the clinical diagnosis of congenital liver disorders, abnormally elevated serum transaminase activities due to unknown causes were associated with a higher rate of a genetic diagnosis (Table 2). We next performed a case-level review to assess concordance between genotype and phenotype. Among 43 liver disease patients with P/LP variants (Supplementary Information 3), we confirmed a previous clinical diagnosis for eleven, identified a genetic disease that partially explained the phenotype for eleven, reclassified disease for seven, identified a molecular subtype of inherited liver diseases for six, and identified a cause for undiagnosed liver diseases for five. We also recommended further workup in three patients to confirm or refute the liver diagnosis (Fig. 3C). In addition, we examined the phenotypic concordance for the 25 kidney patients carrying P/LP variants in liver genes: 15/25 patients had a corresponding liver phenotype, which were mostly attributable to P/LP variants in genes like *PKD1,* MODY or ciliopathy genes causing both kidney and liver disease (Supplement Information 3). Benefits of a genetic diagnosis included the ability to guide familial testing and obtain an early diagnosis of affected family members for 24 families, or to perform surveillance for known complications, such as brain aneurysms in individuals carrying a pathogenic variant in *PKD1*. Four patients with *HFE* and *PFIC2* will be followed

Genotypes	HC (n = 7856)	CKD (n = 2187)	CLD (n = 758)
HFE variants			
Homozygous C282Y	2	1[$]	1[$]
Homozygous H63D	9	3[&]	0
Compound C282Y/H63D*	60	11	2
C282Yor H63D/PTV	22	5	1[$]
Total (n = 117)	93 (1.20%)	20 (0.91%)	4 (0.53%)
SERPINA1 variants			
Pi ZZ	0	0	3[$]
Pi MZ (n = 222)	138 (1.76%)	49 (2.24%)	15 (1.98%)
Pi SS	1	0	0
Pi MS (n = 606)	482 (6.14%)	92 (4.21%)	32 (4.22%)
Pi Z/ Pi S* (n = 12)	7	5	0

Table 1. *HFE* and *SERPINA1* variants in three cohorts. [&]Individuals with a homozygous pathogenic variant, or two heterozygous pathogenic variants reported in HGMD or ClinVar; *Phase was not determined to evaluate if these two variants are *in-trans*, or *cis*. [$]Cases with a sufficient clinical evidence of liver phenotypes consistent with the genetic diagnosis. One case with H63D has liver phenotypes consistent with hemochromatosis.

Characteristics	Total	No genetic Dx (n, %)	Monogenic Dx (n, %)	*P* value
Total number of individuals	758	715 (94.3%)	43 (5.7%)	
Gender				
Male	351	328 (93.4%)	23 (6.6%)	0.415
Female	407	387 (95.1%)	20 (4.9%)	
Age group				
0–21 year	255	232 (91.0%)	23 (9.0%)	0.019
22–44 year	150	141 (94.0%)	9 (6.0%)	
45–64 year	226	218 (96.5%)	8 (3.5%)	
≥ 65 year	127	124 (96.9%)	3 (2.4%)	
Self-declared race/ethnicity				
White	367	349 (95.1%)	18 (4.9%)	0.903
Hispanic	136	128 (94.1%)	8 (5.9%)	
Black	95	88 (92.6%)	7 (7.4%)	
Asia	61	57 (93.4%)	4 (6.6%)	
Other or unspecified	99	93 (93.9%)	6 (6.1%)	
Primary liver diagnosis				
Metabolic/Congenital	17	7 (41.2%)	10 (58.8%)	6.48e-19
NAFLD/NASH	182	174 (95.6%)	8 (4.1%)	
AIH/PBC/PSC	128	122 (85.3%)	6 (4.7%)	
Abnormal LFT	52	47 (90.4%)	5 (9.6%)	
Biliary atresia	76	73 (96.1%)	3 (3.9%)	
Other	303	292 (96.4%)	11 (3.6%)	

Table 2. Clinical characteristics for monogenic diagnoses in the liver cohort from ES analysis. Chi-square test was used to compare the difference between two groups, either negative or positive genetic diagnosis through ES.

clinically for progression to appropriate stages of disease for cancer screening. Patients with *PGM1* and *PHKA2* pathogenic variants, diagnostic of congenital disorders of glycosylation, can benefit from selective nutritional management. Other implications for better treatment include targeted therapy, clinical trials, or surgical options. For example, a review of clinicaltrials.gov identified 255 clinical trials are enrolling patients with monogenic forms of liver diseases identified in this study (Fig. 3D).

Discussion

Our primary goal was to evaluate the utility of ES for diagnosis of liver disease. Currently available clinical genetic testing for heritable liver diseases exists and is mostly utilized in the pediatric populations. For instance, one lab provides a panel of 72 genes for well-defined monogenic liver diseases, especially cholestasis and biliary atresia[26].

To guide the ES analysis, we developed a list of 502 genes associated with a Mendelian disease with potential liver phenotypes (Fig. 4). This work constitutes an initial attempt at a gene list for monogenic liver disease, but the list will have to be continuously annotated and updated to include new information about genes and variants. For example, we updated the list to include several genes (*TULP3*[27], *KIF12*, *USP53*[28], *KCNN3*[29,30], *GIMAP5*[9]) which have been implicated in monogenic disorders associated with liver phenotypes during the performance of this study. We also removed some genes which, in retrospect, did not have a secure causal relationship with CLD. In the future, the creation of a liver disease workgroup, for instance, under the ClinGen platform or PanelApp[31], will accelerate the development of a reference gene list for CLD.

The current challenge of genetic analysis is to determine the pathogenicity of variants. In this work, we focused on genes associated with monogenic disorders and omitted analysis of risk factors, such as *PNPLA3*. Consistent with prior studies of other genetic disorders, our variant level analyses indicated that many previously reported P/LP variants for liver diseases are too common to be pathogenic and are erroneously annotated in reference databases. We report liver disease genes with the most frequently encountered false-positive P/LP variants to help with the reannotation of reference databases (Supplementary Information 3). We also performed a manual annotation of the data, which confirmed that the application of hard filters for allele frequency and sequence quality may lead to the omission of true pathogenic variants (Fig. 4). For example, in addition to the high frequency *CFTR, HFE* and *SERPINA1* pathogenic variants, two patients with progressive familial intrahepatic cholestasis type 3 carried an *ABCB4* Ala934Thr missense variant which has a MAF of 1.2% in African-American populations, and should be interpreted as a pathogenic variant (Supplementary Information 2). Likewise, a pathogenic p.Lys414fs variant in carnitine palmitoyltransferase (CPT) II gene have an allele frequency above 0.1% in Ashkenazi Jewish population[32]. Thus, a balanced disease-specific approach was necessary for maximizing the diagnostic rate. A case-level review indicated that the genetic results were consistent with the clinical findings in the majority of liver and kidney disease cases, validating our approach. The genetic findings had many implications for diagnosis, risk stratification, surveillance, treatment and management, including potential eligibility for clinical trials. For the patients who did not have a concordant liver disease phenotype, the P/LP variants may be non-penetrant, disease may develop in the future, or the variant may be downgraded in the future based on evidence of non-pathogenicity. We note that our study is limited by the lack of clinical information for most self-reported healthy controls, which hampers our ability to determine the causality of P/LP variants in this cohort.

Altogether, our single-center study indicates a significant diagnostic utility for ES in the evaluation of patients with CLD. Currently, the clinical genetic diagnoses are limited by several pitfalls based on ES. First, ES cannot find P/LP variants in intronic regions or poorly covered regions; second, we did not do homozygous CNV

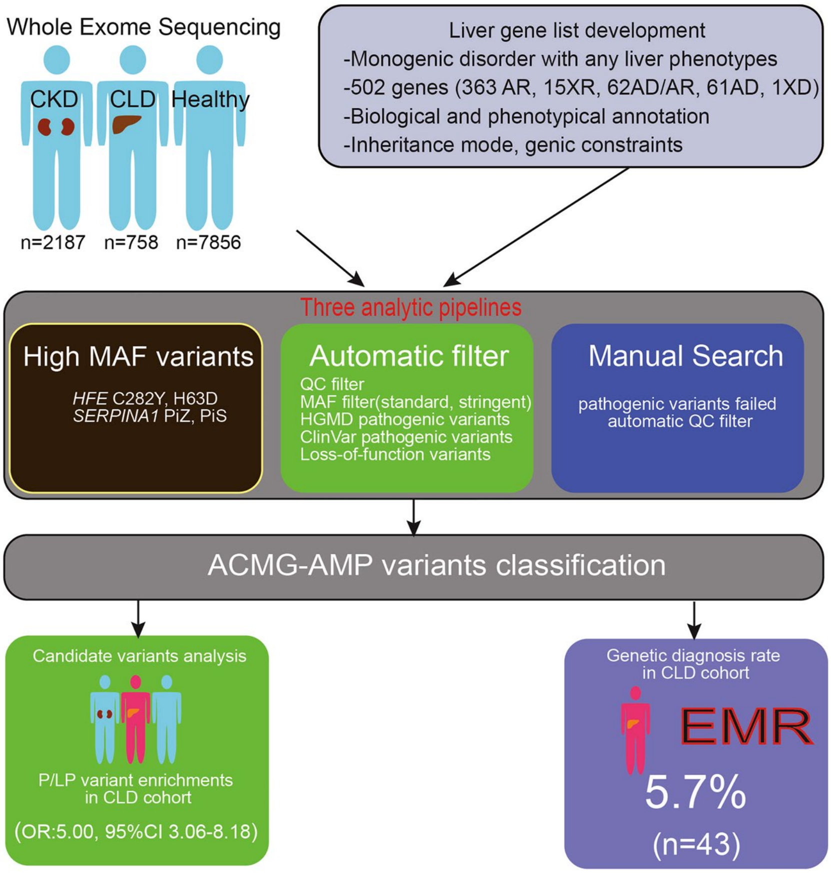
Whole Exome Sequencing
CKD
CLD
Healthy
n=2187
n=758
n=7856
Liver gene list development
-Monogenic disorder with any liver phenotypes
-502 genes (363 AR, 15XR, 62AD/AR, 61AD, 1XD)
-Biological and phenotypical annotation
-Inheritance mode, genic constraints
Three analytic pipelines
High MAF variants
HFE C282Y, H63D
SERPINA1 PiZ, PiS
Automatic filter
QC filter
MAF filter(standard, stringent)
HGMD pathogenic variants
ClinVar pathogenic variants
Loss-of-function variants
Manual Search
pathogenic variants failed automatic QC filter
ACMG-AMP variants classification
Candidate variants analysis
P/LP variant enrichments in CLD cohort
(OR:5.00, 95%CI 3.06-8.18)
Genetic diagnosis rate in CLD cohort
EMR
5.7%
(n=43)

Figure 4. A summary of the genetic analytic strategy and outcomes for liver diseases.

calls for ES data and might miss heterozygous CNVs or small genomic deletions, such as *DNAJB1-PRKACA* in fibrolamellar hepatocellular carcinoma. Third, as we define alternative alleles should be above 30% of total reads from genomic DNA extracted from blood, mosaic and somatic genetic disorders could not be ruled out. Lastly, those with single P/LP heterozygous variants in recessive inheritance gene were excluded for further analysis in this manuscript. Ideally, those patients with a single P/LP heterozygous variant should be identified with further efforts to investigate the corresponding clinical phenotypes. If clinical phenotypes are consistent with a recessive disorder, searching for an additional *in-trans* variant may be important to guide the genetic diagnosis. Therefore, combining WGS and RNA-Seq of liver biopsy may increase the genetic diagnosis rate. Future studies will have to evaluate the diagnostic utility across varied healthcare settings, apply different genetic testing strategies and prospectively demonstrate the impact of genetic testing on clinical decision-making, cost-effectiveness and genetically stratified clinical trials.

Material and methods

Developing a list of monogenic disorders associated with liver phenotypes

We first composed gene list, or "liver gene list", to identify genes causing monogenic diseases with a wide range of liver manifestations. We used Online Mendelian Inheritance in Man database (OMIM), Orphanet, and the Human Phenotype Ontology (HPO) database[33] to search for potential genes with Mendelian inheritance that have been associated with or shown to be causative in liver disease before December 2018. For the search, we used a total of 30 keywords or phrases (Supplementary Fig. 1), then manually reviewed OMIM and related literature[34–39]. We excluded: (1) genes not reported to be linked to any abnormal liver phenotypes; (2) genes within a locus reported from linkage analysis without any known pathogenic variants; (3) genes only discovered in GWAS but lacking any evidence for Mendelian inheritance; (4) genes with only somatic variants reported in abnormal liver phenotypes; (5) genes within a locus associated with abnormal liver phenotypes due to chromosomal abnormalities. The selected genes were annotated for biological functions, clinical liver presentations, and gene constraint score[40]. We annotated the inheritance mode of liver genes based on OMIM and ClinGen then manually curated the list of genes by reviewing relevant literature. The current gene list is an initial attempt to catalog monogenic liver diseases and remains a work in progress. We anticipate that this list will require regular updates and curations and may serve as the basis for a reference liver gene list that can be curated by an expert group, such as ClinGen[41].

Cohorts

We analyzed ES data obtained by sequencing of genomic DNA extracted from peripheral blood of 758 patients with CLD. We enrolled patients from both pediatric and adult liver clinics at Columbia University Irving Medical Center (CUIMC) who were interested in and able to consent to participating in genetics research, without setting inclusion or exclusion criteria (Table 2). In the CLD cohort, 53.7% of participants were female, and 33.6% of participants were under 22 years of age. 182 patients with CLD (24%) were diagnosed with nonalcoholic fatty liver disease (NAFLD) or nonalcoholic steatohepatitis (NASH), 128 patients (16%) with AIH or PBC or PSC, other patients with viral hepatitis (n = 125), and alcoholic hepatitis (n = 27) were also included. A few cases with acute liver failure (n = 5), or hepatocellular carcinoma (HCC, n = 3), or hepatoblastoma (n = 7), or cardiogenic liver cirrhosis (n = 9) were sequenced and analyzed altogether. A selection bias might occur as we attempt to enroll those who might have a genetic cause of liver diseases. The CKD cohort was included because we had access to health records through CUIMC, enabling us to evaluate the penetrance of monogenic liver disorders in a cohort not ascertained for liver disease. Informed consent in writing was obtained from each patient and the study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by CUIMC Institutional Review Board.

Sequence analysis and variant annotation

Sample preparation, target-enrichment, sequencing process, read alignment, and variant calling were previously published[20,21]. We focused on variants that were predicted to have at least moderate to strong biological effects toward protein function and excluded those in intergenic and promoter regions. We used stringent quality filters and removed potential technical false-positive insertions and deletions (indels) using ATAV as previously described[20,42]. We excluded variants failing quality cutoffs in gnomAD or those identified as sequencing artifacts through a comparison of in-house control sequencing data. Current guidelines recommend considering all variants with a minor allele frequency (MAF) of less than 1% at the population level. Thus, we filtered the variants based on the overall MAF of less than 1% in the Genome Aggregation Database (gnomAD)[43]. Variants previously reported as pathogenic were identified using the HGMD and ClinVar. We included only those annotated as pathogenic/likely pathogenic (P/LP) in ClinVar or disease-causing mutation (DM) in HGMD without any conflicting evidence within each database. In addition, we identified novel protein-truncating variants (PTVs) not previously reported in either HGMD or ClinVar. As the initial yield of individuals carrying candidate pathogenic variants was significantly higher than expected, we employed a stringent filter by inheritance mode and subpopulation MAF based on the data from gnomAD and Exome Aggregation Consortium (ExAC): MAF $\leq 10^{-4}$ for dominant disorders and MAF $\leq 10^{-3}$ for recessive disorders[20,23,24]. We used Loss-Of-Function Transcript Effect Estimator (LOFTEE) filter to exclude PTVs with a false prediction. A detailed description of genetic terminology in this study has been described previously[20].

Manual variant classification and clinical data review

Two independent genetic analysts performed a first-tier, stringent analysis of the CLD cohort to reach a consensus classification according to the ACMG-AMP guidelines[44]. We next performed a second-level manual curation of the CLD cohort using lower stringency filters, which identified several well-defined pathogenic variants that were excluded because they either have a MAF above 1% in some ethnic subpopulations or did not pass the stringent sequencing quality filters. This procedure had been successfully used to increase diagnostic yield in prior studies[44,45]. Subsequently, a multidisciplinary group of experts, including genetic counselors, geneticists, molecular pathologists, and clinicians, reviewed the available clinical information in individuals carrying P/LP variants to detect phenotypic concordance with the associated mode of inheritance of disease. If diagnostic evidence was insufficient based on chart review, a follow-up plan was recommended to clarify the significance of the genetic findings.

Statistical analysis

We compared the probability of being loss-of-function intolerant (pLI) and Z scores for genes using an analysis of variance (ANOVA) test to compare differences between the three groups. We analyzed the clinical variables between those with and those without a genetic diagnosis using the Chi-squared test. All statistics and genetic analyses were done in R statistical software (Version 4.0.0). A p-value of <0.05 was considered significant after correction for multiple hypothesis testing.

Ethics declaration statement

I attest that the research included in this report was conducted in a manner consistent with the principles of research ethics, such as those described in the Declaration of Helsinki and/or the Belmont Report. In particular, this research was conducted with the voluntary, informed consent of all research participants, free of coercion or coercive circumstances, and received Columbia University Irving Medical Center Institutional Review Board (IRB) approval consistent with the principles of research ethics and the legal requirements of the lead authors' jurisdictions.

Data availability

The datasets generated and analyzed during the current study are available in the ClinVar repository: https://www.ncbi.nlm.nih.gov/clinvar/. Accession IDs: SCV004024075-SCV004024150.

Received: 20 March 2023; Accepted: 6 September 2023
Published online: 06 December 2023

References

1. QuickStats. Death Rates* for Chronic Liver Disease and Cirrhosis,† by Sex and Age Group—National Vital Statistics System, United States, 2000 and 2015. *MMWR Morb. Mortal. Wkly. Rep.* **66**(38), 1031. https://doi.org/10.15585/mmwr.mm6638a9 (2017).
2. Barton, J. C., Edwards, C. Q. & Acton, R. T. HFE gene: Structure, function, mutations, and associated iron abnormalities. *Gene* **574**(2), 179–192. https://doi.org/10.1016/j.gene.2015.10.009 (2015).
3. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13**(4), 399–408. https://doi.org/10.1038/ng0896-399 (1996).
4. Honeyman, J. N. *et al.* Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science (80-)* **343**(6174), 1010–1014. https://doi.org/10.1126/science.1249484 (2014).
5. Belkaya, S. *et al.* Inherited IL-18BP deficiency in human fulminant viral hepatitis. *J. Exp. Med.* **216**(8), 1777–1790. https://doi.org/10.1084/jem.20190669 (2019).
6. Haack, T. B. *et al.* Biallelic mutations in NBAS cause recurrent acute liver failure with onset in infancy. *Am. J. Hum. Genet.* **97**(1), 163–169. https://doi.org/10.1016/j.ajhg.2015.05.009 (2015).
7. Cousin, M. A. *et al.* RINT1 bi-allelic variations cause infantile-onset recurrent acute liver failure and skeletal abnormalities. *Am. J. Hum. Genet.* https://doi.org/10.1016/J.AJHG.2019.05.011 (2019).
8. Vilarinho, S. *et al.* Recurrent recessive mutation in deoxyguanosine kinase causes idiopathic noncirrhotic portal hypertension. *Hepatology* **63**(6), 1977–1986. https://doi.org/10.1002/hep.28499 (2016).
9. Drzewiecki, K. *et al.* GIMAP5 maintains liver endothelial cell homeostasis and prevents portal hypertension. *J. Exp. Med.* https://doi.org/10.1084/JEM.20201745 (2021).
10. Denny, J. C. *et al.* The "All of Us" research program. *N. Engl. J. Med.* **381**(7), 668–676. https://doi.org/10.1056/NEJMsr1809937 (2019).
11. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science (80-)* https://doi.org/10.1126/science.aaf6814 (2016).
12. Kong, X. F. *et al.* Identification of a ferrochelatase mutation in a Chinese family with erythropoietic protoporphyria. *J. Hepatol.* **48**(2), 375–379. https://doi.org/10.1016/j.jhep.2007.09.013 (2008).
13. Li, X.-H. *et al.* Clinical and molecular characterization of Wilson's disease in China: Identification of 14 novel mutations. *BMC Med. Genet.* **12**(1), 6. https://doi.org/10.1186/1471-2350-12-6 (2011).
14. Kong, X.-F. *et al.* Recurrent porphyria attacks in a Chinese patient with a heterozygous PBGD mutation. *Gene* **524**(2), 401–402. https://doi.org/10.1016/j.gene.2013.03.130 (2013).
15. Karlsen, T. H., Lammert, F. & Thompson, R. J. Genetics of liver disease: From pathophysiology to clinical practice. *J. Hepatol.* **62**(S1), S6–S14. https://doi.org/10.1016/j.jhep.2015.02.025 (2015).
16. Pelusi, S. *et al.* Clinical exome sequencing for diagnosing severe cryptogenic liver disease in adults: A case series. *Liver Int.* **42**(4), 864–870. https://doi.org/10.1111/LIV.15185 (2022).
17. Liebe, R. *et al.* Diagnosis and management of secondary causes of steatohepatitis. *J. Hepatol.* **74**, 1455–1471. https://doi.org/10.1016/j.jhep.2021.01.045 (2021).
18. Masserot-Lureau, C. *et al.* Incidence of liver abnormalities in Fanconi anemia patients. *Am. J. Hematol.* **87**(5), 547–549. https://doi.org/10.1002/ajh.23153 (2012).
19. Sharma, D. *et al.* Tip of the iceberg: A comprehensive review of liver disease in Inborn errors of immunity. *Hepatology* **76**(6), 1845–1861. https://doi.org/10.1002/HEP.32539 (2022).
20. Rasouly, H. M. *et al.* The burden of Candidate pathogenic variants for kidney and genitourinary disorders emerging from exome sequencing. *Ann. Intern. Med.* **170**(1), 11–21. https://doi.org/10.7326/M18-1241 (2019).
21. Groopman, E. E. *et al.* Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**(2), 142–151. https://doi.org/10.1056/nejmoa1806891 (2019).

22. Milo Rasouly, H. *et al.* The burden of candidate pathogenic variants for kidney and genitourinary disorders emerging from exome sequencing. *Ann. Intern. Med.* https://doi.org/10.7326/M18-1241 (2019).
23. Whiffin, N. *et al.* Using high-resolution variant frequencies empowers clinical genome interpretation and enables investigation of genetic architecture. *Am. J. Hum. Genet.* **104**(1), 187–190. https://doi.org/10.1016/J.AJHG.2018.11.012 (2019).
24. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**(10), 1151–1158. https://doi.org/10.1038/GIM.2017.26 (2017).
25. Shashi, V. *et al.* A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. *Genet. Med.* **21**(1), 161–172. https://doi.org/10.1038/S41436-018-0044-2 (2019).
26. Syed, A. & Hajira, A. A comprehensive review of progressive familial intrahepatic cholestasis (PFIC): Genetic disorders of hepatocanalicular transporters. *Gastroenterol. Res.* https://doi.org/10.14740/gr609e (2014).
27. Devane, J. *et al.* Progressive liver, kidney, and heart degeneration in children and adults affected by TULP3 mutations. *Am. J. Hum. Genet.* **109**(5), 928–943. https://doi.org/10.1016/J.AJHG.2022.03.015 (2022).
28. Maddirevula, S. *et al.* Identification of novel loci for pediatric cholestatic liver disease defined by KIF12, PPM1F, USP53, LSR, and WDR83OS pathogenic variants. *Genet. Med.* **21**(5), 1164–1172. https://doi.org/10.1038/S41436-018-0288-X (2019).
29. Koot, B. G. P., Alders, M., Verheij, J., Beuers, U. & Cobben, J. M. A de novo mutation in KCNN3 associated with autosomal dominant idiopathic non-cirrhotic portal hypertension. *J. Hepatol.* **64**(4), 974–977. https://doi.org/10.1016/J.JHEP.2015.11.027 (2016).
30. Bauer, C. K. *et al.* Gain-of-function mutations in KCNN3 encoding the small-conductance Ca 2þ-activated K þ channel SK3 cause Zimmermann–Laband syndrome. *Am. J. Hum. Genet.* **104**, 1139–1157. https://doi.org/10.1016/j.ajhg.2019.04.012 (2019).
31. Stark, Z. *et al.* Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. *Am. J. Hum. Genet.* **108**(9), 1551–1557. https://doi.org/10.1016/J.AJHG.2021.06.020 (2021).
32. Taggart, R. T., Smail, D., Apolito, C. & Vladutiu, G. D. Novel mutations associated with carnitine palmitoyltransferase II deficiency. *Hum. Mutat.* **13**(3), 210–220. https://doi.org/10.1002/(SICI)1098-1004(1999)13:3%3c210::AID-HUMU5%3e3.0.CO;2-0 (1999).
33. Köhler, S. *et al.* The human phenotype ontology in 2017. *Nucleic Acids Res.* **45**(D1), D865–D876. https://doi.org/10.1093/nar/gkw1039 (2017).
34. Marques-da-Silva, D. *et al.* Liver involvement in congenital disorders of glycosylation (CDG). A systematic review of the literature. *J. Inherit. Metab. Dis.* **40**(2), 195–207. https://doi.org/10.1007/s10545-016-0012-4 (2017).
35. Kulecka, M. *et al.* Clinical applicability of whole-exome sequencing exemplified by a study in young adults with the advanced cryptogenic cholestatic liver diseases. *Gastroenterol. Res. Pract.* **2017**, 1–8. https://doi.org/10.1155/2017/4761962 (2017).
36. Schonfeld, E. A. & Brown, R. S. Genetic testing in liver disease: What to order, in whom, and when. *Clin. Liver Dis.* **21**(4), 673–686. https://doi.org/10.1016/j.cld.2017.06.001 (2017).
37. Stephens, M. C., Boardman, L. A. & Lazaridis, K. N. Individualized medicine in gastroenterology and hepatology. *Mayo Clin. Proc.* **92**(5), 810–825. https://doi.org/10.1016/j.mayocp.2017.03.002 (2017).
38. Nicastro, E. & D'Antiga, L. Next generation sequencing in pediatric hepatology and liver transplantation. *Liver Transplant.* **24**(2), 282–293. https://doi.org/10.1002/lt.24964 (2018).
39. Pericleous, M., Kelly, C., Ala, A. & Schilsky, M. L. The epidemiology of rare hereditary metabolic liver diseases. In *Clinical Epidemiology of Chronic Liver Diseases* 307–330 (Springer, 2018). https://doi.org/10.1007/978-3-319-94355-8_17.
40. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**(7616), 285–291. https://doi.org/10.1038/nature19057 (2016).
41. Rehm, H. L. *et al.* ClinGen: The clinical genome resource. *N. Engl. J. Med.* **372**(23), 2235–2242. https://doi.org/10.1056/NEJMSR1406261 (2015).
42. Ren, Z., Povysil, G. & Goldstein, D. B. ATAV: A comprehensive platform for population-scale genomic analyses. *bioRxiv* https://doi.org/10.1101/2020.06.08.136507 (2020).
43. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7 (2020).
44. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**(5), 405–424. https://doi.org/10.1038/gim.2015.30 (2015).
45. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**(2), 249–255. https://doi.org/10.1038/gim.2016.190 (2017).

Acknowledgements

Thank all the participants for enrolling into first studies. The study was supported by the Columbia precision medicine initiative.

Author contributions

X.F.K., A.G.G., D.B.G., and J.W. were responsible for conception and design, gene list searching and curation, analysis and interpretation of the data, and manuscript writing. X.F.K., K.B., E.E.G., P.R.S., A.T.W., A.G.G., and J.W. contributed to variants annotation and clinical review of affected individuals. S.K., J.Z., M.M., J.M.V., L.M.D., M.M., E.C.V., H.J.W., T.C.W. and J.W. were involved in patient recruitment, administrative, technical, or logistical support. E.E.G., A.T.W., E.C., H.M.R., B.Z., and S.S. coordinated the gene list search and curation.

Funding

Dr. Xiao-Fei Kong was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K08DK128631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Dr. Gharavi has received research funding from the Renal Research Institute and Natera and has served as a consultant for Goldfinch Bio. All other authors declare no conflicts of interest.

Competing interests

Dr. Xiao-Fei Kong was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K08DK128631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Dr. Gharavi has received research funding from the Renal Research Institute and Natera and has served as a consultant for Goldfinch Bio. All other authors declare no conflicts of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-42202-1.

Correspondence and requests for materials should be addressed to X.-F.K. or J.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.