Exome Sequencing in Individuals with Isolated Biliary Atresia

Ramakrishnan Rajagopalan[1], Ellen A. Tsai[1,2,3], Christopher M. Grochowski[1,4], Susan M. Kelly[5], Kathleen M. Loomes[6,7], Nancy B. Spinner[1,8], Marcella Devoto[7,9,10]*

[1]Division of Genomic Diagnostics, Dept. of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[2]Genomics and Computational Biology Graduate Group, The University of Pennsylvania, Philadelphia PA, USA.

[3]Genetic Epidemiology Group, Dept. of Translational Biology, Biogen, Cambridge, MA USA.

[4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.

[5]Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA.

[6]Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[7]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[8]Dept. of Pathology and Laboratory Medicine, The Perelman School of Medicine at The University of Pennsylvania, Philadelphia PA.

[9]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[10]Department of Translational and Precision Medicine, University La Sapienza, Rome, Italy.

*devoto@email.chop.edu

Abstract

Biliary atresia (BA) is a severe pediatric liver disease resulting in necroinflammatory obliteration of the extrahepatic biliary tree. BA presents within the first few months of life as either an isolated finding or with additional syndromic features. The etiology of isolated BA is unknown, with evidence for infectious, environmental, and genetic risk factors described. However, to date, there are no definitive causal genes identified for isolated BA in humans, and the question of whether single gene defects play a major role remains open. We performed exome-sequencing in 100 North American patients of European descent with isolated BA (including 30 parent-child trios) and considered several experimental designs to identify potentially deleterious protein-altering variants that may be involved in the disease. In a case-only analysis, we did not identify genes with variants shared among more than two probands, and burden tests of rare variants using a case-case control design did not yield significant results. In the trio analysis of 30 simplex families (patient and parent trios), we identified 66 *de novo* variants in 66 genes including a nonsense variant, p.(Cys30Ter), in the gene *STIP1*. *STIP1* is a co-chaperone for the heat-shock protein, *HSP90AA1*, and has been shown to be a modifier of a biliary phenotype in zebrafish exposed to a known BA-associated toxin (biliatresone). *Conclusion*: Our results do not support the hypothesis that a simple genetic model is responsible for the majority of cases of isolated BA. Our finding of a *de novo* mutation in a candidate gene for BA (*STIP1*) that confers sensitivity to the toxin biliatresone in zebrafish suggests further exploration of how genetic susceptibility and environmental exposure interact to cause BA is warranted.

Introduction

Biliary atresia (BA) is an obstructive cholangiopathy with initial symptoms arising during the first days to weeks of life. BA occurs as an isolated finding in 85% of affected individuals, and with additional syndromic features (heterotaxy and/or other congenital birth defects) in 15% [1]. The incidence of BA varies across different populations, with estimates ranging from 1 in 5,000 to 1 in 14,000 live births [2]. BA presents clinically with neonatal cholestasis, elevated bilirubin and liver enzymes, although the differential diagnosis suggested by these findings is extensive. The diagnosis of BA is suggested by features of biliary obstruction on liver histology, including bile duct proliferation, portal tract expansion and bile plugs. The diagnosis is confirmed upon intraoperative cholangiogram showing lack of patency of the extrahepatic biliary tree.

Although the etiology of BA is not clear, a genetic component is supported by multiple lines of evidence. These include the presence of familial cases [3-5], case reports of individuals with syndromic BA with likely causal genes identified (including *FOXA2* [6], *CFC1* [7,8], *ZEB2* [9], *ZIC3* [10], *HNF1B* [11] and *PKD1L1* [12]), animal models of BA involving the genes *Sox17* [13], *Foxm1* [14], *Invs* [15], *Onecut1* [16], and candidate susceptibility genes including *ADD3* [17], *GPC1* [18], *EFEMP1* [19] and *ARF6* [20] identified by genome-wide association studies (GWAS). Genes implicated in syndromic BA function in bile-duct development and the establishment of left-right symmetry. A follow-up study of *GPC1* using a zebrafish model confirmed that suppressing *GPC1* expression gives rise to biliary defects[21]. A GWAS identified SNPs associated with BA upstream of *XPNPEP1* and *ADD3* in a Chinese cohort of isolated BA patients [17], and this association has been confirmed in both a Thai [22] and an European-American cohort [23]. In a study of 499

isolated and syndromic BA patients and 1928 controls, our group identified a GWAS signal within 2p16.1 implicating common variants in the gene *EFEMP1* [19]. Another genome-wide association study of 63 BA patients and 1907 controls identified two SNPs downstream of the gene *ARF6* associated with isolated BA [20], although we could not replicate this finding in our study. However, to date, no genes have been definitively identified as a cause of isolated BA. Environmental exposure to virus or toxins has long been proposed as a hypothesis for perinatal form of BA[2]. But the genetic susceptibility factors linking such exposure and BA are yet to be identified.

In this work, we hypothesized that rare and novel, deleterious, protein-altering variants transmitted in a Mendelian fashion are involved in the etiology of isolated BA, and patients with isolated BA are likely to be enriched with such variants. To test this hypothesis, we considered three experimental designs (*Table 1*). First, we conducted a case-only analysis where we looked for potentially deleterious novel variants. In the second design, we conducted a case-control analysis comparing the cumulative frequency of rare variants in the two groups. Finally, we performed trio analysis on a subset of 30 probands to identify genes with rare, protein-altering *de novo* and compound heterozygous or homozygous variants.

Patients and Methods

Patients with isolated BA and controls

We studied a cohort of 100 self-reported white, non-Hispanic patients with isolated BA. The 100 patients were chosen from a larger cohort of 1063 BA cases based on their self-reported race and ethnicity and availability of parental DNA samples. Ninety-nine participants were enrolled in the NIDDK-funded Childhood Liver Disease Research

Network (ChiLDReN) at one of the sixteen participating institutions under Institutional Review Board (IRB) approved protocols (IRB 04-003655 and 06-004814). A set of dizygotic twins were consented into the approved IRB study at the Children's Memorial Hospital, Chicago, IL and only one of them were part of the cohort as the scope of the current work was to analyze unrelated individuals. Informed consent in wirting was obtained from each patient enrolled in this study. The diagnosis of BA was made by clinical presentation, liver histology, and an intraoperative cholangiogram. Most patients also had BA confirmed by examination of the biliary remnant from a Roux-en-Y hepatic portoenterostomy (also known as a Kasai operation). DNA samples were extracted from whole blood or lymphoblastoid cell lines. For a subset of patients (n = 30), DNA samples were available from both parents.

Controls were individuals recruited from two different studies at the Children's Hospital of Philadelphia, PediSeq (the CHOP cohort of the Clinical Sequencing and Exploratory Research (CSER) consortium[24], and VEO-IBD (Very-Early Onset Inflammatory Bowel Disease) [25], and enrolled in institution-approved IRB protocols (IRB 12-009169; 14-010826). Controls from the PediSeq dataset (n=158) fall into four different disease cohorts (cardiovascular, hearing loss, mitochondrial and intellectual disability) but have no known liver disease. Controls from the VEO-IBD cohort (n=145) were the unaffected parents of individuals with VEO-IBD. The patient recruitment, methods, and experimental protocols used in these studies were performed in accordance with relevant guidelines and regulations set by the aforementioned institutions. This study was approved by the ethics committee at the Children's Hospital of Philadelphia.

Exome sequencing (ES)

Genomic DNA (3-5 ug) was supplied to Perkin Elmer, Inc. (Perkin-Elmer, Branford, CT) for whole exome sequencing of the BA patients and their parents. The DNA was amplified and enriched using the Agilent SureSelect V4+UTR 71Mb All Exon Capture Kit (Agilent Technologies, Santa Clara, CA). Sequencing was performed on the Illumina HiSeq 2000 (Illumina, San Diego, CA) at the Perkin-Elmer sequencing service center. Control samples were sequenced using Agilent SureSelect V4 51Mb Exon Capture Kit (Agilent Technologies, Santa Clara, CA) on the Illumina HiSeq 2000 (Illumina, San Diego, CA) at the internal sequencing core at CHOP (BGI@CHOP, Philadelphia, PA). All of the exome sequencing data was produced with paired-end reads of 100 base-pairs in length (2 x 100bp). The unaligned FASTQ reads of all samples were returned to us for read analysis and downstream processing as detailed below.

Variant calling and annotation

Raw sequencing reads were mapped to the GRCh37.69 reference genome using the Burrows-Wheeler Aligner [26] to produce an aligned BAM file. BAM files were realigned around known indels using IndelRealigner in the GenomeAnalysisToolKit software (GATK) and base quality scores were recalibrated using the BaseQualityScoreRecalibrator in GATK [27]. Variant calling was performed using HaplotypeCaller and jointly genotyped using the GenotypeGVCFs command in GATK. The initial variant quality filtering was done using the VariantQualityScoreRecalibrator (VQSR) in GATK. SNPEff [28] was used to annotate the variants with their expected functional consequence. Other annotations such as allele frequencies in non-Finnish Europeans (NFE) from the gnomAD variant database, and deleteriousness scores such as CADD (Combined Annotation Dependent Depletion version 1.4) [29] were added using the

Gemini software [30]. We manually reviewed all variants found in the *de novo* analysis, and variant clusters with overlapping genomic positions that could be indicative of technical artifacts [31].

Population stratification, relatedness and sex check

The Peddy [32] software tool was used to infer genetic ancestry from the exome sequencing data by comparing our samples against the samples from the 1000 Genomes Project Phase 3 [33]. All samples from BA patients and their parents clustered with the European populations in the 1000 Genomes Project. We also confirmed the sex of the participants based on heterozygosity of X chromosome variants. Pairwise relatedness (estimated by the kinship coefficient) between samples was investigated using the KING [34] software and confirmed that there were no related individuals in the entire cohort beside the known parent-child pairs used in the trio analysis.

Variant prioritization

We filtered the variants identified from the ES for stop gain, stop loss, frameshift and in-frame insertions/deletions (InDels), missense and splicing variants. We excluded variants overlapping with segmental duplications and variants with call rate less than 90% from all analyses. In the case-only analysis, we included novel and potentially deleterious variants, defined as variants that were not present in the Non-Finnish European (NFE) population of the gnomAD genomes and exomes and had a phred-scaled CADD score of at least 30, representing the top 0.1% of the potentially deterious variants. For the burden tests in the case-control experimental design, all protein-altering variants with a cohort (cases and controls combined) minor allele frequency (MAF) of <5% were included. We did not apply a CADD score threshold for this analysis. For the *de novo* analysis, we

considered all variants irrespective of their frequency or the CADD score. For the analysis under the recessive model, we retained variants with an allele frequency in gnomAD NFE less than or equal to 0.1% and a CADD score of at least 30. Details pertaining to variant prioritization are provided in *Table 1*.

Autosomal copy-number variant (CNV) analysis

Mean depth of coverage for each individual exonic interval in the autosomes was computed using the DepthOfCoverage routine in the GATK package and copy number variants were called using XHMM following the protocol described elsewhere [35]. We excluded CNVs with quality score (Q_SOME) less than 60 to retain only high quality CNVs. To identify novel CNVs, we excluded any variant found in the Database of Genomic Variants (DGV) [36] and the ExAC [37]. We did not look for CNVs in the sex chromosomes as the ExAC dataset only had calls from autosomes.

Gene, pathway and gene-set definitions

For all the analyses, we used refFlat_hg19 gene definitions from the UCSC genome browser[38]. We used pathway definitions including KEGG (n = 293), Panther (n = 112), Reactome (n = 1530), WikiPathways (n = 437) and BioCarta (n = 237), and annotation gene-sets from Gene Ontology (GO) Biological Process (n = 5192), Molecular Functions (n = 1136) and Cellular Components (n=641) downloaded from the Enrichr website[39].

Case-only and trio analyses: Over representation tests

To determine if our lists of candidate genes from the case-only and trio analyses were enriched for specific gene-sets or pathways, we performed over-representation analysis using the R package enrichR [39]. This software essentially counts the number of genes

with a given annotation that are present in our lists of candidate genes and genome-wide, and compares it to the number of genes that are not associated with the same annotation in our lists and genome-wide, using a Fisher exact test. Over representation tests were carried out on gene-sets and pathways.

Case-control analysis: Rare variant burden tests

To test if a gene, gene-set or pathway contained significantly more cases with at least one rare variant (MAF < 5%) compared to controls, we performed burden tests using the Combined and Multivariate Collapsing (CMC) approach [40] implemented in the exactCMC procedure from the software package rvtests [41]. Burden tests were carried out at different levels including 1) the gene level; 2) the gene-set level (e.g. collection of genes with similar function or structure or cellular localization, such as sodium channel genes or membrane bound receptors); and 3) the pathway level, which includes genes that function in a specific biological pathway (e.g. WNT or NOTCH signaling). The number of cases and controls with and without at least one rare variant in a given gene, gene-set, or pathway were compared using a Fisher exact test.

Trio analysis

Trio analysis was performed using Gemini software[30] in a subset of 30 probands for whom DNA from both parents was available for ES. *De novo* variants were identified using the command `*de_novo*` which essentially identifies SNVs and indels occuring in proband only and not in the parents. Further, we filtered for exonic, protein-altering and splice-site variants. Compound heterozygous variants were identified using the command `*comp_het*` which identifies two heterozygous variants in the same gene inherited from

different parents. Homozygous variants were identified using the `*autosomal_recessive*` command which identifies rare variants transmitted to the proband from both parents.

To test if the total number of *de novo* protein-altering (missense and loss-of-function) variants identified in our cohort was significantly larger than expected, we used the R-package *denovolyzeR*[42]. The *denovolyzeByClass* function implemented in the denovolyzeR package computes the significance of the observed number of exonic variants compared to the pre-computed *de novo* mutation rates for all classes of exonic variants.

Statistical significance

For gene level burden tests, the raw p-values were corrected for multiple testing using a Bonferroni correction for the total number of genes tested (n=16,393). For the pathway based analyses, the raw P-values were corrected for multiple testing using a Bonferroni correction for the total number of pathway definitions (n=2609). For the gene-set based analyses, we corrected for the total number of gene-sets used in the analysis (n = 6969). We considered a test to be statistically significant if the Bonferroni-adjusted p-value was less than or equal to 0.05.

Results

Case-only analysis

In the case-only analysis we identified a total of 496,654 potentially deleterious novel variants in the 100 probands. These included variants that passed the quality filters, of which 63,473 single nucleotide variants (SNV) or small InDels were found to be protein-altering changes (stop gain, stop loss, frameshift and in-frame InDels, missense and

splicing variants). Of these, 332 variants in 326 genes were novel (not in the NFE population in gnomAD exomes and genomes) and had a phred-scaled CADD score of 30 or more. This formed our variant dataset for the case-only analyses. Almost all of the variants (331/332) were present in heterozygosity in a single individual, and 96 out of the 100 individuals in the cohort had at least one potentially deleterious novel variant. Six genes had two different variants in two different individuals *(MAST2, WDR35, DNAH5, FBN2, SCUBE2 and PITPNM3).* Over-representation analysis did not yield any significant gene, pathway or gene-set after correcting for multiple testing. Copy-number variant analysis did not reveal any novel variant in the probands after filtering for variants present in the population databases ExAC and DGV. Details of the SNVs and InDels with annotations are provided in the *Supp. Table 1*.

Case-control analysis

In the second experimental design, we conducted burden tests for rare variants (MAF < 5%) by comparing their cumulative number in a gene, gene-set or pathway in the 100 cases against the 303 controls. After correcting for the multiple testing, we did not identify any gene, gene-set or a pathway with a statistically significant excess of rare variants in the cases compared to the controls.

Trio analysis

In the final experimental design, we sequenced the parents of a subset of the patients (n = 30) to identify *de novo* variants in the probands under the hypothesis that they are more likely to be involved in the etiology of the disease. We identified a total of 66 *de novo*, protein-altering and splice variants in 66 different genes in 25 out of 30 individuals, with a mean of 2.2 variants per individual (range: 0-7; median 3). We did not identify any *de*

novo variants in five individuals. All the *de novo* variants were manually verified in the BAM file and a small subset of them were validated using Sanger sequencing (n = 4). There were 58 missense, 3 frameshift indels, 3 nonsense, and 2 splice-site variants. A prioritized list of 14 *de novo* variants are provided in *Table 2* and the list of all *de novo* variants are provided in the *Supp. Table 2*. The number of protein-altering *de novo* variants identified in this cohort was significantly larger than the expected number of such variants in a cohort of 30 individuals (expected = 21.5; observed = 66; p-value = $1.13x10^{-14}$). This signal was mainly driven by the missense variants (expected = 18.9; observed = 58; p-value = $4.15e^{-13}$), while the loss-of-function variants had a modest significance (expected = 2.6; observed = 8; p-value = $5.61x10^{-3}$). Of the eight putative *de novo* loss-of-function variants, a premature stop-gain in the gene *STIP1* and a splice-acceptor variant in the gene *REV1* were the most interesting as they are known to be depleted for and intolerant to loss-of-function in the general population with a pLI score of 1. The pLI or probability of loss-of-function intolerance is a gene-level measure estimated by comparing the expected number of loss-of-function variants against the observed in a population. The pLI score ranges from 0 to 1 with the highest value being the most intolerant. Over-representation analysis of the 66 genes did not yield significant results after correcting for multiple testing. We did not identify any compound heterozygous or homozygous variants at a maximum alternate allele frequency threshold of 0.1% and a phred-scaled CADD score of 30.

Discussion

In this study, exome sequencing was used to identify genomic changes in a cohort of patients with isolated biliary atresia, and in spite of the utilization of several distinct

strategies, we were unable to demonstrate an associated genetic change. We did identify a candidate gene by analysis of 30 trios, finding a *de novo* likely pathogenic variant in *STIP1*, which we hypothesize may result in susceptibility to toxin induced biliary disease.

BA presents as an isolated finding in 85% of cases, and in a syndromic form with laterality and/or other congenital malformations in 15%. Mouse models of BA (*Sox17, Onecut1*) and gene mutations in patients with syndromic forms of BA (*CFC1, ZIC3, FOXA2, PKD1L1*) suggest an etiology involving genetic factors that influence left-right symmetry, and bile duct development and morphology. Genome-wide association studies in patients with isolated BA have suggested candidate genes and genomic regions (*ADD3, XPNaaPEP1* in 10q25.1, *EFEMP1* in 2p16.1, *GPC1* in 2q37.3, *ARF6* in 14q21.3) but consistent proof of the involvement of each of these regions has been elusive. Biliary atresia as a consequence of exposure to environmental toxins is also proposed as a model for the perinatal form of the disease based on research into the natural occurance of BA in Australian sheep, which led to the discovery of an isoflavonoid (Biliatresone) found in specific plant species, later found to cause biliary defects in zebrafish [43]. RNA-seq experiments on Biliatresone treated and control fish identified several differentially expressed stress signaling pathways including the Glutathione antioxidant pathway, cellular redox response, and heat shock response [44]. However, the genetic susceptibility factors that might link exposure to these agents and biliary atresia are yet to be identified. In summary, from previous literature, there are genes that are known to cause syndromic forms of BA in humans and mice, genes that are associated with isolated BA identified via GWAS, and biological pathways that are perturbed in zebrafish treated with the toxin Biliatresone, which has been associated with

biliary atresia in sheep. However, in spite of these clues, there has not been a causal gene identified for isolated BA in humans.

In this work, we performed exome sequencing in 100 North American unrelated patients of European ancestry with isolated BA and looked for novel or rare protein-altering variants that may explain the BA phenotype (strategies presented in *Table 1*). Such study designs have been successful in finding novel genes responsible for rare Mendelian diseases [45]. Fewer exome sequencing studies with sufficiently large sample sizes have been successfully employed in identifying associations with complex phenotypes [46,47]. This study was primarily designed with the goal of finding genes with rare and highly deleterious variants that may explain the occurrence of isolated BA. We considered three experimental designs including a case-only analysis to look for novel, potentially deleterious and protein-altering variants, a case-control analysis to look for an excessive burden of rare variants in the cases compared to the controls, and a trio analysis, to look for *de novo* and homozygous or compound heterozygous variants.

In the case-only design, we identified 332 novel and potentially deleterious protein-altering variants in our cohort of 100 individuals with isolated BA. The majority of the variants were found in a single individual in a heterozygous state. Six genes had variants in two individuals, and the variants were different from each other. This suggests that BA is unlikely to result from a single genetic change, although we cannot rule out extreme locus heterogeneity, with a small percentage of cases having a genetic etiology, perhaps in combination with an environmental insult. Case-control analysis did not yield any statistically significant result and this may suggest that there are no variants or genes with

large enough effect sizes present in the cases as compared to the controls or simply the sample size is too small to identify such a difference between cases and controls.

Trio-based exome sequence analysis identified 66 *de nov*o protein-altering variants in 66 genes in 25 out of 30 patients, a number significantly higher than expected (p-value = 4.15×10^{-13}). All putative loss-of-function *de novo* variants and a prioritized list of *de novo* missense variants with a CADD score greater than or equal to 30 are provided in *Table 2*. Of all the genes with putative loss-of-function *de novo* variants, a splice-acceptor variant in the gene *REV1* and a premature stop in one of the transcripts in the gene *STIP1* were the most interesting for several reasons. These genes seem to be intolerant for loss-of-function variants with a perfect pLI score of 1. *STIP1* is a chaperone that assists in the transfer of proteins from *HSP70* to *HSP90*, which are an integral part of the heat shock response pathway and as mentioned below, *HSP90* down-regulation has been implicated in BA livers. Mass spectrometry experiments to identify differentially expressed proteins in 20 BA liver biopsies compared to 12 non-BA, neonatal cholestasis livers found that heat shock protein 90 (*HSP90*) was significantly down-regulated in BA livers, and was the most significantly altered protein. This suggests that *HSP90* could serve as a potential biomarker for BA [48]. We therefore hypothesized that a gene-environment interaction may explain isolated BA where a predisposing genetic factor exists and an external insult triggers the initiation of the disease. Based on the *de novo* variant found in the human exome data, the identification of the heatshock response pathway as differentially expressed in zebrafish, our collaborators performed a CRISPR/Cas9 based experiment to introduce a frameshift mutation in exon 1 of the zebrafish *stip1* gene to create heterozygous mutant fish. When treated with toxin, *stip1* heterozygous fish were

highly sensitive to a low dose of biliatresone compared to the wild-type [49]. *REV1* encodes a protein similar to the *S. cerevisiae* mutagenesis protein *Rev1* and is known to be involved in DNA repair. REV1 is shown to be regulated by the heat-shock protein HSP90 in the DNA repair pathway [50] and further functional studies are warranted to understand its role in BA.

Our study has several limitations including sample size (which is a function of the disease frequency) and the failure of exome sequencing to identify all possible disease causing genomic changes. The sample size was relatively small for statistical analyses such as rare variant burden tests and for the trio analysis whose aim was to identify genes with recurring *de novo* variants. Exome sequencing is a targeted capture experiment which focuses only on the exonic regions, and coverage might not be uniform through the coding parts of the genome. This limits our ability to study intronic regions and/or poorly targeted parts of the genome. Copy-number detection from exome sequencing is challenging and XHMM software tool used in this work is tuned for specificity rather than sensitivity. Finally, exome sequencing does not allow us to look for structural variation in these samples.

Together, our analysis of rare and novel coding variants does not support a simple genetic model in which a small number of genes are responsible for the majority of cases with isolated BA. However, the identification of a loss-of-function *de novo* variant in the gene *STIP1* and the existing evidence for its sensitivity to the biliatresone toxin in zebrafish opened new opportunities for further investigating BA as a consequence of environmental exposure. Potentially, other genes in the heat shock response pathway or other pathways identified in the zebrafish experiments could be considered as candidates.

Acknowledgments

We would like to thank all the patients and their families who participated in this study; the ChiLDReN investigators, research coordinators and the Data Coordinating Center for the ChiLDReN network. We would like to acknowledge Dr. Judith Kelsen and Dr. Ian Krantz for providing exome sequencing data for the control samples. We also would like to acknowledge Dr. David Piccoli and Dr. Barbara Haber for their support. We would like to acknowledge Dr. Michael Pack and Dr. Xiao Zhao at the University of Pennsylvania for the collaborative work on zebrafish models.

Author Contribudions

MD, KL, and NBS designed the experiments. RR and ET performed the computational and bioinformatics analysis. CG performed validation studies. KL and SK recruited patients. RR, ET, KL, NBS and MD were involved in the interpretation of the results. RR wrote the manuscript and all authors contributed in editing the manuscript.

Financial Support

This work is funded by the NIH grant R01-DK090045 and funds from the Fred and Suzanne Biesecker Pediatric Liver Center at Children's Hospital of Philadelphia. The Childhood Liver Disease Research Network is supported by U01 grants from the National Institute of Diabetes, Digestive and Kidney Disease (NIDDK): DK062481, DK062456, DK062497, DK084536, DK062500, DK062503, DK062466, DK062453, DK062452, DK062436, DK103149, DK103135, DK084575, DK084538, DK062470, DK103140 and DK062455. In addition, the project was supported by UL1 grants from the National Institutes of Health Clinical and Translational Sciences Award (CTSA) program through the National Center for Advancing Translational Science (NCATS):

TR000130, TR001872, TR001857, TR001108, TR002535, TR000454, TR000423, RR025014, TR000077 and TR000003.

Competing interests

The authors declare no competing interests.

List of Tables

Table 1: Experimental designs and variant prioritization schemes used in this study

Table 2: Prioritized list of *de novo* variants identified in the trio analysis of 30 parent-child trios. This table lists all the loss-of-function, splice variants and missense variants with a CADD score greater than 30.

Supporting Information

Supplementary Table 1: List of all the *de novo* variants identified

Supplementary Table 2: List of all the novel variants identified in the 100 probands

References

1 Schwarz, K. B. *et al.* Extrahepatic anomalies in infants with biliary atresia: results of a large prospective North American multicenter study. *Hepatology (Baltimore, Md.)* **58**, 1724-1731, doi:10.1002/hep.26512 (2013).

2 Lakshminarayanan, B. & Davenport, M. Biliary atresia: A comprehensive review. *J Autoimmun* **73**, 1-9, doi:10.1016/j.jaut.2016.06.005 (2016).

3 Gunasekaran, T. S., Hassall, E. G., Steinbrecher, U. P. & Yong, S. L. Recurrence of extrahepatic biliary atresia in two half sibs. *American journal of medical genetics* **43**, 592-594, doi:10.1002/ajmg.1320430317 (1992).

4 Strickland, A. D., Shannon, K. & Coln, C. D. Biliary atresia in two sets of twins. *The Journal of pediatrics* **107**, 418-420 (1985).

5 Lachaux, A. *et al*. Familial extrahepatic biliary atresia. *Journal of pediatric gastroenterology and nutrition* **7**, 280-283 (1988).

6 Tsai, E. A. *et al.* Heterozygous deletion of FOXA2 segregates with disease in a family with heterotaxy, panhypopituitarism, and biliary atresia. *Human mutation* **36**, 631-637, doi:10.1002/humu.22786 (2015).

7 Davit-Spraul, A., Baussan, C., Hermeziu, B., Bernard, O. & Jacquemin, E. CFC1 gene involvement in biliary atresia with polysplenia syndrome. *Journal of pediatric gastroenterology and nutrition* **46**, 111-112, doi:10.1097/01.mpg.0000304465.60788.f4 (2008).

8 Jacquemin, E., Cresteil, D., Raynaud, N. & Hadchouel, M. CFCI gene mutation and biliary atresia with polysplenia syndrome. *Journal of pediatric gastroenterology and nutrition* **34**, 326-327 (2002).

9 Cui, S., Erlichman, J., Russo, P., Haber, B. A. & Matthews, R. P. Intrahepatic biliary anomalies in a patient with Mowat-Wilson syndrome uncover a role for the zinc finger homeobox gene zfhx1b in vertebrate biliary development. *Journal of pediatric gastroenterology and nutrition* **52**, 339-344, doi:10.1097/MPG.0b013e3181ff2e5b (2011).

10 Ware, S. M. *et al.* Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. *American journal of human genetics* **74**, 93-105, doi:10.1086/380998 (2004).

11 Kotalova, R. *et al.* Hepatic phenotypes of HNF1B gene mutations: a case of neonatal cholestasis requiring portoenterostomy and literature review. *World J Gastroenterol* **21**, 2550-2557, doi:10.3748/wjg.v21.i8.2550 (2015).

12 Berauer JP, M. A., Okoul D, Sabo A, Hegde M, Chopra P, Perlmutter DH, Bull L, Thompson, RL, Spinner N, Loomes KM, Rajagopalan R, Guthery SL, Moore B, Shneider BL, Kamath BM, Molleston JP, Bezerra JA, Murray KF, Alonso E, Rosenthal P, Squires RH, Sokol RJ, Wang KS, Magee JC, Sherker AH, Karpen SJ. Identification of PKD1L1 Gene Variants in Biliary Atresia Patients with Structural Malformations (BASM). *AASLD 2017* (2017).

13 Uemura, M. *et al.* Sox17 haploinsufficiency results in perinatal biliary atresia and hepatitis in C57BL/6 background mice. *Development (Cambridge, England)* **140**, 639-648, doi:10.1242/dev.086702 (2013).

14 Krupczak-Hollis, K. *et al.* The mouse Forkhead Box m1 transcription factor is essential for hepatoblast mitosis and development of intrahepatic bile ducts and

vessels during liver morphogenesis. *Dev Biol* **276**, 74-88, doi:10.1016/j.ydbio.2004.08.022 (2004).

15 Shimadera, S. *et al.* The inv mouse as an experimental model of biliary atresia. *Journal of pediatric surgery* **42**, 1555-1560, doi:10.1016/j.jpedsurg.2007.04.018 (2007).

16 Clotman, F. *et al.* The onecut transcription factor HNF6 is required for normal development of the biliary tract. *Development (Cambridge, England)* **129**, 1819-1828 (2002).

17 Garcia-Barcelo, M. M. *et al.* Genome-wide association study identifies a susceptibility locus for biliary atresia on 10q24.2. *Human molecular genetics* **19**, 2917-2925, doi:10.1093/hmg/ddq196 (2010).

18 Leyva-Vega, M. *et al.* Genomic alterations in biliary atresia suggest region of potential disease susceptibility in 2q37.3. *Am J Med Genet A* **152A**, 886-895, doi:10.1002/ajmg.a.33332 (2010).

19 Chen, Y. *et al.* A genome-wide association study identifies a susceptibility locus for biliary atresia on 2p16.1 within the gene EFEMP1. *PLoS Genet* **14**, e1007532, doi:10.1371/journal.pgen.1007532 (2018).

20 Ningappa, M. *et al.* The Role of ARF6 in Biliary Atresia. *PloS one* **10**, e0138381, doi:10.1371/journal.pone.0138381 (2015).

21 Cui, S. *et al.* Evidence from human and zebrafish that GPC1 is a biliary atresia susceptibility gene. *Gastroenterology* **144**, 1107-1115.e1103, doi:10.1053/j.gastro.2013.01.022 (2013).

22 Kaewkiattiyot, S., Honsawek, S., Vejchapipat, P., Chongsrisawat, V. & Poovorawan, Y. Association of X-prolyl aminopeptidase 1 rs17095355 polymorphism with biliary atresia in Thai children. *Hepatol Res* **41**, 1249-1252, doi:10.1111/j.1872-034X.2011.00870.x (2011).

23 Tsai, E. A. *et al.* Replication of a GWAS signal in a Caucasian population implicates ADD3 in susceptibility to biliary atresia. *Human genetics* **133**, 235-243, doi:10.1007/s00439-013-1368-2 (2014).

24 Green, R. C. *et al.* Clinical Sequencing Exploratory Research Consortium: Accelerating Evidence-Based Practice of Genomic Medicine. *American journal of human genetics* **99**, 246, doi:10.1016/j.ajhg.2016.06.002 (2016).

25 Kelsen, J. R. *et al.* Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease. *Gastroenterology* **149**, 1415-1424, doi:10.1053/j.gastro.2015.07.006 (2015).

26 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

27 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics* **43**, 11.10.11-33, doi:10.1002/0471250953.bi1110s43 (2013).

28 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).

29 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).
30 Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS computational biology* **9**, e1003153, doi:10.1371/journal.pcbi.1003153 (2013).
31 Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)* **30**, 2843-2851, doi:10.1093/bioinformatics/btu356 (2014).
32 Pedersen, B. S. & Quinlan, A. R. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *American journal of human genetics* **100**, 406-413, doi:10.1016/j.ajhg.2017.01.017 (2017).
33 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
34 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).
35 Fromer, M. & Purcell, S. M. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr Protoc Hum Genet* **81**, 7 23 21-21, doi:10.1002/0471142905.hg0723s81 (2014).
36 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic acids research* **42**, D986-992, doi:10.1093/nar/gkt958 (2014).
37 Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* **48**, 1107-1111, doi:10.1038/ng.3638 (2016).
38 Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic acids research* **46**, D762-D769, doi:10.1093/nar/gkx1020 (2018).
39 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).
40 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics* **83**, 311-321, doi:10.1016/j.ajhg.2008.06.024 (2008).
41 Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics (Oxford, England)* **32**, 1423-1426, doi:10.1093/bioinformatics/btw079 (2016).
42 Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Curr Protoc Hum Genet* **87**, 7 25 21-15, doi:10.1002/0471142905.hg0725s87 (2015).
43 Lorent, K. *et al.* Identification of a plant isoflavonoid that causes biliary atresia. *Science translational medicine* **7**, 286ra267, doi:10.1126/scitranslmed.aaa1652 (2015).
44 Zhao, X. *et al.* Glutathione antioxidant pathway activity and reserve determine toxicity and specificity of the biliary toxin biliatresone in zebrafish. *Hepatology (Baltimore, Md.)* **64**, 894-907, doi:10.1002/hep.28603 (2016).

45 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-755, doi:10.1038/nrg3031 (2011).

46 Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102-106, doi:10.1038/nature13917 (2015).

47 Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *American journal of human genetics* **94**, 233-245, doi:10.1016/j.ajhg.2014.01.010 (2014).

48 Dong, R. *et al.* Identification of HSP90 as potential biomarker of biliary atresia using two-dimensional electrophoresis and mass spectrometry. *PloS one* **8**, e68602, doi:10.1371/journal.pone.0068602 (2013).

49 Xiao Zhao, R. R., Kathleen M. Loomes, Nancy B. Spinner, Michael Pack. Heat shock chaperone STIP1 as a modulator and genetic risk factor of extra-hepatic biliary atresia [Abstract]. *Hepatology (Baltimore, Md.)* **66 (1 Suppl)**, 59A (2017).

50 Pozo, F. M. *et al.* Molecular chaperone Hsp90 regulates REV1-mediated mutagenesis. *Mol Cell Biol* **31**, 3396-3409, doi:10.1128/MCB.05117-11 (2011).

Table 1: Experimental designs used in this study

Experimental design	**Case-only**	**Case-control**	**Trio**	
			de novo	**Recessive**
Subjects	100 cases	100 cases, 303 controls	30 parent-child trios	30 parent-child trios
CADD score threshold	≥ 30	none	none	≥ 30
AF threshold in gnomAD NFE	0	none	none	≤ 0.1%
AF threshold in cohort	none	< 5%	none	none
Statistical test	Over-representation	Burden	Over-representation	NA

* Variant types include stop gain, stop loss, frameshift and in-frame InDels, missense, splicing (splicing variants include both canonical sites (+/- 1,2 as well as 3-8 bp into the introns)).

**** Abbreviations in Table 1**

CADD	Combined Annotation-Dependent Depletion score
AF	Allele Frequency
gnomAD	Genome Aggregation Database
NFE	Non-Finnish Europeans

Table 2: Prioritized list of *de novo* variants identified in the trio analysis of 30 parent-child trios. This table lists all the loss-of-function, splice variants and missense variants with a CADD score greater than 30.

Chromosome	Position	Reference allele	Alternate allele	dbSNP ID	Gene symbol	Impact	Amino acid change
All *de novo* loss-of-function variants							
2	100038117	T	C		REV1	splice_acceptor	NM_001037872:exon11:c.1674-2A>G NM_016316:exon11:c.1677-2A>G
3	47048735	CTG	C		NBEAL2	frameshift	NM_015175:exon47:c.7230_7231del:p.T2410fs
8	145621814	CCT	C	rs782640869	CPSF1	frameshift	NM_013291:exon25:c.2823_2824del:p.S941fs
9	114804380	C	A	rs781608602	SUSD1	splice_donor	NM_001282640:exon17:c.2214+1G>T
11	63953386	C	A	rs545010478	STIP1	stop_gained	NM_001282652:exon1:c.C90A:p.C30X
17	34165489	C	CTT		TAF15	frameshift	NM_003487:exon11:c.836_837insTT:p.P279fs NM_139215:exon11:c.845_846insTT:p.P282fs
20	16360632	A	T		KIF16B	stop_gained	NM_001199865:exon19:c.T2015A:p.L672X NM_001199866:exon19:c.T2015A:p.L672X NM_024704:exon19:c.T2015A:p.L672X
X	140969336	C	A		MAGEC3	stop_gained	NM_138702:exon4:c.C663A:p.Y221X
Non-synonymous *de novo* variants w/ CADD $\geq$ 30							
2	63834044	C	T	.	MDH1	missense	NM_001199112:exon8:c.C661T:p.R221C NM_001199111:exon9:c.C982T:p.R328C NM_005917:exon9:c.C928T:p.R310C
2	68873329	C	T	rs771784209	PROKR1	missense	NM_138964:exon1:c.C376T:p.R126C
4	158142842	G	T	.	GRIA2	missense	NM_000826:exon2:c.G112T:p.D38Y NM_001083619:exon2:c.G112T:p.D38Y
5	52780025	G	A	.	FST	missense	NM_006350:exon4:c.G623A:p.G208E NM_013409:exon4:c.G623A:p.G208E
11	4095877	C	T	.	STIM1	missense	NM_001277961:exon7:c.C937T:p.R313C NM_001277962:exon7:c.C937T:p.R313C NM_003156:exon7:c.C937T:p.R313C
17	38319096	G	A	.	CASC3	missense	NM_007359:exon6:c.G727A:p.D243N

**** Abbreviations in Table 2**

dbSNP	Database of Single Nucleotide Polymorphisms
CADD	Combined Annotation-Dependent Depletion score