Manuscript Number: GASTRO 18-01393

Title: A Missense Variant in *PTPN22* is a Risk Factor for Drug-induced Liver Injury

Elizabeth T. Cirulli[1]*, Paola Nicoletti[2,3*], Karen Abramson[4], Raul J. Andrade[5], Einar S. Bjornsson[6], Naga Chalasani[7], Robert J. Fontana[8], Pär Hallberg[9], Yi Ju Li [4,10], M Isabel Lucena[5], Nanye Long[11], Mariam Molokhia[12], Matthew R. Nelson[13], Joseph A. Odin[14], Munir Pirmohamed[15], Thorunn Rafnar[16], Jose Serrano[17], Kari Stefansson[16], Andrew Stolz[18], Ann K. Daly[19], Guruprasad P. Aithal[20#] and Paul B. Watkins[21,22#]

on behalf of Drug-Induced Liver Injury Network (DILIN) investigators and International DILI consortium (iDILIC)

[1] Duke Center for applied Genomics and Precision Medicine, Duke University, Durham, North Carolina;
[2] Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, New York;
[3] Sema4, a Mount Sinai venture, Stamford, Connecticut, USA
[4]Duke Molecular Physiology Institute, Duke University, Durham, North Carolina
[5]UGC Digestivo, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospital Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain; Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd);
[6] Department of Internal Medicine, Landspitali University Hospital, Reykjavik, Iceland;
[7] Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, Indiana;
[8] University of Michigan, Ann Arbor, Michigan;
[9] Department of Medical Sciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden;
[10]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina
[11] Institute for Cyber-enabled Research, Michigan State University, East Lansing, Michigan
[12]School of Population Health & Environmental Sciences, King's College, London, UK;
[13] Target Sciences, GSK, King of Prussia, Pennsylvania;
[14] Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York;
[15]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK;
[16] deCODE genetics, 101 Reykjavik, Iceland
[17]National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland;
[18] University of Southern California, Los Angeles, California;
[19] Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK;
[20] Nottingham Digestive Diseases Centre and National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre at the Nottingham University Hospital NHS Trust and University of Nottingham, Nottingham, UK;
[21] UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina
[22] University of North Carolina Institute for Drug Safety Sciences, RTP, North Carolina

*Authors share co-first authorship; #Authors share co-senior authorship

Funding

The DILIN (https://dilin.dcri.duke.edu/) is supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) as a Cooperative Agreement (U01s) under grants: U01-DK065176 (Duke), U01-DK065201 (UNC), U01-DK065184 (Michigan), U01-DK065211 (Indiana), U01DK065193 (UConn), U01-DK065238 (UCSF/CPMC), U01-DK083023 (UTSW), U01-DK083027 (TJH/UPenn), U01-DK082992 (Mayo), U01-DK083020 (USC), U01-DK100928 (Icahn). Additional funding is provided by CTSA grants UL1 RR025761 (Indiana), UL1 RR025747 (UNC), and UL1 UL1 RR024986 (UMich). The iDILIC study was supported by the International Serious Adverse Events Consortium which received funding from Abbott, Amgen, Daiichi-Sankyo, GlaxoSmithKline, Merck, Novartis, Pfizer, Roche, Sanofi-Aventis, Takeda, and the Wellcome Trust. DILIGEN and iDILIC sample collection was funded by the National Institute for Health Research (NIHR) Nottingham Digestive Diseases Biomedical Research Unit at the Nottingham University Hospitals NHS Trust and University of Nottingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. GPA is the gastrointestinal and liver disorder theme lead for the NIHR Nottingham BRC (Reference no: BRC-1215-20003). The EUDRAGENE collaboration received support from the EC 5th Framework program (QLRI-CT-2002-02757). The Spanish DILI Registry is partly funded by the Spanish Medicine Agency, Fondo Europeo de Desarrollo Regional - FEDER (FIS PI16_01748, PI15 _01440). CIBERehd is funded by Instituto de Salud Carlos III. The Swedish case collection (SWEDEGENE) has received support from the Swedish Medical Products Agency, the Swedish Society of Medicine (2008-21619), Swedish Research Council (Medicine 521-2011-2440), and Swedish Heart and Lung Foundation (20120557). MM was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London.

Abbreviations: DILI (drug-induced liver injury); genome wide association study (GWAS); Odd Ratio (OR); Roussel Uclaf Causality Assessment Method (RUCAM); Major Histocompatibility Complex (MHC); allele frequency (AF); Human leukocyte antigen (HLA); Single Nucleotide Polymorphism (SNP); Amoxicillin-clavulanate (AC)

Correspondence to: Paola Nicoletti, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai One Gustave Levy Place, New York, NY, USA 10029. Email: paola.nicoletti@mssm.edu

Conflicts of interest: The authors disclose the following: Dr. Nelson is an employee of GlaxoSmithKline. Dr. Nicoletti is an employee of Sema4. Drs. Chalasani, Fontana, and Watkins report consulting agreements and research grants with several pharmaceutical companies but none represent potential conflicts for this paper. Dr. Rafnar and Dr. Stefansson are employees of deCODE genetics/Amgen. The DILIN causality committee considers potential conflicts while assigning cases for adjudication to individual investigators. The remaining authors disclose no conflicts.

Author contribution:

Study concept and design: Guruprasad P. Aithal, Paul B. Watkins, Elizabeth T. Cirulli, Paola Nicoletti, and Ann K. Daly

Case recruitment and data acquisition: Paul B. Watkins, Naga Chalasani, Robert J. Fontana, Jose Serrano, Andrew Stolz, Huiman X. Barnhart, ,Guruprasad P. Aithal, Einar Bjornsson, Thorunn Rafnar, Kari Stefansson, Raul J. Andrade, Pär Hallberg, M. Isabel Lucena, Mariam Molokhia, Joseph A. Odin, Munir Pirmohamed, and Ann K. Daly

Case adjudication: Paul B. Watkins Robert J. Fontana, Andrew Stolz, Naga Chalasani, Joseph A. Odin, Jose Serrano, and the DILIN Investigators, Guruprasad P. Aithal and Einar Bjornsson for iDILIC

Sample preparation and laboratory analysis: Karen Abramson

Data analysis and interpretation: Elizabeth T. Cirulli, Paola Nicoletti, Yi-Ju Li, Paul B. Watkins, Guruprasad P. Aithal, Thorunn Rafnar, Kari Stefansson, and Ann K. Daly

Writing the manuscript: Elizabeth T. Cirulli, Paola Nicoletti, Paul B. Watkins, Guruprasad P. Aithal, Ann K. Daly, Naga Chalasani, and Robert Fontana

Abstract:

Background & Aims: We performed genetic analyses of a multiethnic cohort of patients with idiosyncratic drug-induced liver injury (DILI) to identify variants associated with susceptibility.

Methods: We performed a genome-wide association study of 2048 individuals with DILI (cases) and 12,429 individuals without (controls). Our analysis included subjects of European (1806 cases and 10,397 controls), African American (133 cases and 1,314 controls), and Hispanic (109 cases and 718 controls) ancestry. We analyzed DNA from 113 Icelandic cases and 239,304 controls to validate our findings.

Results: We associated idiosyncratic DILI with rs2476601, a nonsynonymous polymorphism that encodes a substitution of tryptophan with arginine in the protein tyrosine phosphatase, non-receptor type 22 gene (*PTPN22*) (odds ratio [OR], 1.44; 95% CI, 1.28–1.62; $P=1.2\text{x}10^{-9}$ and replicated the finding in the validation set (OR, 1.48; 95% CI, 1.09–1.99; $P=.01$). The minor allele frequency showed the same effect size (OR > 1) among ethnic groups. The strongest association was with amoxicillin and clavulanate-associated DILI in persons of European ancestry (OR, 1.62; 95% CI, 1.32–1.98; $P=4.0\text{x}10^{-6}$; allele frequency=13.3%), but the polymorphism was associated with DILI of other causes (OR, 1.37; 95% CI, 1.21–1.56; $P=1.5\text{x}10^{-6}$; allele frequency=11.5%). Among amoxicillin- and clavulanate-associated cases of European ancestry, rs2476601 doubled the risk for DILI among those with the HLA risk alleles A*02:01 and DRB1*15:01.

Conclusions: In a genome-wide association study, we identified rs2476601 in *PTPN22* as a non-HLA variant that associates with risk of liver injury caused by multiple drugs and validated our finding in a separate cohort. This variant has been associated with increased risk of autoimmune diseases, providing support for the concept that alterations in immune regulation contribute to idiosyncratic DILI.

KEY WORDS: amino acid change; GWAS; mutation; inflammation

Introduction

Idiosyncratic drug-induced liver injury (DILI) is a rare adverse drug reaction that is an important cause of acute liver failure in the developed world[1, 2]. DILI typically occurs in 10 to 20 out of 100,000 treated patients, and while it can lead to death, most cases resolve with discontinuation of the offending drug[3, 4]. DILI is nonetheless one of the most frequent complications in the development and approval of new drugs, often leading to failure in the later stages of drug development or regulatory actions, including post-marketing withdrawals[5].

A number of genome-wide association studies (GWAS) on DILI have been performed, leading to the discovery of significant associations with several HLA alleles that are generally drug-specific. For example, HLA-B*57:01 is associated with DILI in response to flucloxacillin, HLA-A*02:01 and HLA-DRB1*15:01 are associated with amoxicillin-clavulanate (AC), HLA-B*35:02 is associated with minocycline, and HLA-A*33:01 is associated with terbinafine and probably several other drugs as well [6-9]. The association of DILI risk with HLA alleles supports a role for adaptive immunity in DILI. Although no confirmed associations outside the HLA region have yet been identified in a GWAS[10], a trend toward association with all cause DILI was recently observed with a SNP (rs2476601) in the PTPN22 gene [7]. Because this variant has been associated with risk for a variety of autoimmune diseases, confirming this association would provide further support for the immune basis for DILI.

The Drug Induced Liver Injury Network (DILIN) in collaboration with the International Drug Induced Liver Injury Consortium (iDILIC), has assembled a cohort of 2,048 DILI cases and 12,429 population controls across three major ethnic populations (Europeans, African-Americans and Hispanics). After conducting a trans-ethic meta-analysis, we replicated our top associated SNPs on an independent European cohort of cases and performed multiple subset analyses to investigate their relationship with known HLA risk alleles for DILI and their effect sizes within a certain drug or injury type. Here, we confirm a significant association of DILI risk with rs2476601 in the PTPN22 gene. This is the first GWAS significant association outside the HLA region and the first that appears to hold across many different classes of drugs. Our finding supports immune mechanisms having a broad role in DILI.

Materials and Methods

We carried out a case/control association study in three different populations (Europeans, Hispanics and African-Americans) and then performed a meta-analysis.

Cases

In the current study, we analyzed 2048 DILI cases due to multiple causal drugs, including amoxicillin-clavulonate (AC) and flucloxacillin, collected by the iDILIC and DILIN consortia. Causality assessment was performed as previously reported[8, 11]. 1149 European cases were previously genotyped and analyzed by Urban et al.[11] and/or Nicoletti et al.[8] and 899 had undergone GWAS genotyping for the first time. Table S1 shows the breakdown of the case cohorts by recruitment center, genotyping chip and ethnicity. Clinical characteristics of the DILI subjects were reported in Table 1.

DILIN Cases

A total of 1074 DILIN cases were included, of which 443 DILIN Caucasian cases had been previously reported[8, 11], and 631 cases, consisting of 389 European, 133 African-American and 109 Hispanic descendants were newly genotyped[12]. The DILIN protocol and entrance criteria have been previously published[12]. All participants provided written informed consent. Causality assessment was performed as previously described, and only cases considered probably, highly likely, or definitely related to the implicated drug were included[13]. DNA was extracted from lymphocytes and stored at the NIDDK biosample repository at Rutgers University, Piscataway, NJ. Genome-wide genotyping for the 564 DILIN cases was performed with the Multi Ethnic Genome Illumina Array at Duke University and for 32 African Americans and 35 Hispanic DILIN cases was performed with the 1Million Illumina duo Array at Duke University.

iDILIC European Cases

A total of 974 European iDILIC cases with a range of causal drugs and recruitment phases were included. Of these, 706 cases had been described previously,[8, 11] while 268 cases due to flucloxacillin or AC were newly genotyped. The 268 patients were recruited between May 2009

and May 2013 as a part of an international collaborative study involving international recruitment centers. All participants provided written informed consent, and each study had been approved by the appropriate national or institutional ethical review boards. The clinical inclusion criteria for all cases were those described by Aithal et al.[14] The iDILIC cases were evaluated by application of the Council for International Organizations of Medical Science (CIOMS) scale, also called the Roussel Uclaf Causality Assessment Method (RUCAM),[15] and by expert review by a panel of three hepatologists. Only cases having at least possible causality (score ≥3) were included in the study. DNA was prepared as described previously[6]. Genome-wide genotyping of the additional iDILIC cases was performed by the Broad Institute, Boston by Infinium HumanCoreExome BeadChip for 167 cases and by Infinium Human OmniExpress BeadChip for 101 cases.

Clinical characteristics of the DILI cases

We collected additional clinical information to further investigate the relevance of the most significant associations. Time from start of medication to DILI recognition, concomitant medications, and maximum serum levels of alkaline phosphatase (ALP) and (ALT) were available for both DILIN and iDILIC cases. Also available on all cases was whether the injury was hepatocellular, cholestatic or mixed (based on the initial R value[15] . Specific diagnosis of autoimmune diseases was recorded for iDILIC cases, while DILIN cases included reports on whether patients had a general history of autoimmune/collagen vascular disease (Yes/No).

Controls

As DILI has a very low incidence, we consider that a large set of general population control samples can overcome the potential bias of inclusion of people who may have experienced unrecorded DILI events. 10,397 European controls described in our previous study[8] were used. Moreover, data for 1,314 African American and 718 Hispanic controls were obtained from the MESA study (phs000420.v6.p3) in dbGAP.[16] Table S1 also shows the breakdown of the control cohorts by genotyping chip and ethnicity.

Genetic analysis

Quality control (QC) checks on the initial genotype data were performed as summarized in Supplemental Materials and Methods. EIGENSTRAT analysis was used to identify the Caucasian, African-American and Hispanic cohorts. Our final sample sizes were 1,806 cases and 10,397 controls of European ancestry, 133 cases and 1,314 controls of African-American ancestry, and 109 cases and 718 controls of Hispanic ancestry. SNP imputation was performed in batches dividing the samples according to ethnicity and genotyping platforms. For each batch imputation was carried out using Michigan Imputation Server[17]. as described in the Supplementary Materials and Methods. For HLA genotypes, four digit HLA alleles were inferred using HIBAG[18].

Association analyses were performed using logistic regression under the additive model in Plink [19]. EIGENSTRAT axes were used as covariates. A meta-analysis of the three ethnic groups was performed using GWAMA fixed effects. Variants that 1) had a concordant effect in at least two of the ethnic groups and that 2) showed final meta p-values below $5x10^{-8}$ were considered statistically significant[20, 21]. The top associated imputed SNPs were genotyped in the available cases using a TaqMan® SNP genotyping assay (ThermoFisher Scientific, Waltham, MA) in accordance with the manufacturer's recommendations.

Multi-marker association analysis among the combinations of carriage groups of known HLA risk alleles and the PTPN22 variant allele was performed by logistic regression using principal component as covariates and considering the joint negative carriers as the reference group in the drug-specific cohorts (such as AC, Flucloxacillin, Terbinafine and Flupirtine). Moreover, after transforming the quantitative clinical variables (latency, maximum ALP and maximum ALT) to improve normality, we applied a linear regression model to test differences among known HLA risk alleles and PTPN22 variant in AC cohort.

Epistasis analysis was performed by logistic regression, using principal component axes as covariates and considering an interaction term between being a carrier of any HLA risk alleles and being a carrier of the associated variant.

We also performed association analysis between the PTPN22 variant and reported clinical variables. First, we treated latency and other clinical variables as a quantitative trait. After transforming the quantitative variables to improve normality, we applied a linear regression model to test PTPN22 variant effect on clinical trait in a cases-only design. Epistasis, multi-markers and multinomial logistic regression analyses were carried out using STATA15.

Icelandic DILI replication cohort

An independent Icelandic DILI replication cohort was recruited at the National University Hospital of Iceland. The Icelandic DILI cases were evaluated in accordance with iDILIC causality assessment criteria[8]. Clinical characteristics of the Icelandic DILI subjects are reported in Table S2. The cohort included 113 DILI cases and 239,304 population controls. Geneotyping data from the icelandic sample set was imputed as previously described [22,23, 24]. HLA alleles were also imputed by Graphtyper[25]. Logistic regression under an additive model was used to test for association between variants and DILI. Detailed description of Icelandic analyses is reported in the Supplementary Materials and Methods.

Results

Overall findings

Our final meta-analysis included 3,622,749 SNPs in 2,048 cases and 12,429 controls (see QQ plots in Figure S1). Clinical characteristics of the well-phenotyped DILI cases across three main ethnicities are reported in Table 1. We identified a significant association with rs2476601 (chr1:114377568>A/G), a polymorphism changing tryptophan to arginine at codon 620 of *PTPN22* (OR 1.44 95%CI [1.28-1.62] $P=1.2x10^{-9}$; Figure 1). The enrichment was observed across all ethnic groups analyzed in our study, although the low number of African-American and Hispanic cases limited the power to identify a significant association for variants with an OR < 2 (Table 2). Similar odds ratios were also evident within subgroups of European ancestry

(Table S3). Independent genotyping of available DILIN cases across ethnicities (N=1070) confirmed the GWAS genotypes for rs2476601 with a concordance rate of 100% (Table S4). rs2476601 was also found to increase the risk of DILI in the independent Icelandic replication cohort (AF = 0.13 in 113 cases vs 0.09 in 239,304 controls), having an effect size that was comparable to that of the discovery cohort (OR=1.48, 95%CI [1.09-1.99] P= 0.01).

In addition to rs2476601, several variants in the MHC region were found to have genome-wide significant p-values, led by rs3129880 (OR=1.48 95%CI [1.36-1.60] $P=1.2x10^{-20}$). This variant is a proxy of HLA-DRB1*15:01 ($r^2 = 0.56$) consistent with the large number of AC DILI cases in the cohort. As expected based on inclusion of 195 flucloxacillin and 444 AC Caucasian DILI cases, the most significant HLA risk alleles were HLA-B*57:01, followed by HLA-DRB1*15:01 (OR = 2.19, $P = 1.4x10^{-18}$, see Table 2). Unlike the rs2476601 association, HLA association signals were specific for the European population in which flucloxacillin and AC cases were the most abundant.

Subsequent genome-wide conditional analysis incorporating the genotypes of the four well-established DILI HLA risk alleles (HLA-B*57:01, HLA-DRB1*15:01, HLA-A*02:01 and HLA-A*33:01)[6-8] as covariates was undertaken to identify novel independent risk factors. The analysis revealed that rs2476601 remained the most significant independent risk variant (OR=1.45 95%CI [1.30-1.64] $P = 7.6x10^{-10}$, Figure 1B and Table S5). Similarly, the independence between rs2476601 and the main HLA risk alleles was confirmed in the Icelandic cohort in a multivariate regression model (OR=1.54; P = 0.013, Table S6). When controlling for the four major known DILI HLA risk alleles, HLA-C*04:01 was the most significant independent HLA allele associated with DILI risk reaching near statistical significance when corrected for the total number of imputed HLA alleles (OR=1.21; 95%CI [1.09-1.37]; $P = 6.3x10^{-4}$). HLA-C*04:01 association showed consistent trends across all three ethnicity groups (European P=0.004, OR=1.19; African-American P=0.02, OR=1.42; Hispanic P=0.53, OR=1.13, Table S7). Data for individual drugs in relation to this risk association are shown in Table S8. It is notable that the greatest association was seen with the 58 cases where DILI was attributed to herbal and dietary supplements (OR 2.24, p = 0.0008, individual agents listed in Table S9).

We also found that rs72631546, an intergenic marker on chromosome 2, was the third most significant variant (OR = 1.84, P = $1.2x10^{-7}$, Table 2). rs72631546 is in LD (r^2 = 0.5) with rs72631567, which was a SNP previously suspected to be associated with DILI risk [8]. The rs72631546 association was consistent between Europeans and Hispanics (OR = 1.79 P = $1.9x10^{-6}$; OR = 2.07 P = 0.003 respectively) and was independent of the known HLA risk associations (OR = 1.87 P = $6.0x10^{-8}$).

Association with PTPN22 rs2476601

In the European cohort, the AC cases showed the most significant association with rs2476601 (OR=1.62 95%CI [1.32-1.98] P = $4.0x10^{-6}$) with higher frequency than European controls (AF = 0.13 vs AF = 0.08, respectively). We also found evidence that the association was consistent among the remaining of European DILI cases (N=1362 OR= 1.37 95%CI [1.21-1.56] P= $1.5x10^{-6}$, AF= 0.11, Figure S2), and it did not appear to be driven by particular drugs or categories of drugs (Table 3). Significance of $P \leq 0.05$ was seen for cases due to several causal drugs including sulfamethoxazole-trimethoprim (P= 0.01) and terbinafine (P= 0.01). On the other hand, cases related to drugs such as flucloxacillin and diclofenac, which were well represented in the cohort as causal agents, showed smaller increases in minor allele frequency (AF) compared with controls, which were not statistically significant ($P > 0.05$). We also evaluated the relationship of rs2476601 genotype to DILI phenotype. Of our European ancestry cases, 45% had a hepatocellular pattern of injury, and 55% were cholestatic or mixed. We found enrichment compared to controls for rs2476601 in both injury patterns, with similar AF and ORs (hepatocellular AF = 0.12, OR=1.38, P = 0.0001; cholestatic/mixed AF = 0.13, OR = 1.50, P = $6.5x10^{-8}$; Table S10). We also found that there was a trend for the frequency of rs2476601 to be higher in the DILI cases most confidently ascribed to the implicated drug (Figure S2). We found that there was no significant association between rs2476601 and time of onset of DILI relative to starting treatment with the implicated drug as well as maximum ALP or maximum ALT values.

Assessment of correlation with autoimmune diseases

As rs2476601 has been previously associated with numerous autoimmune diseases[26] we investigated whether the presence of autoimmune diseases in our DILI cohort could have contributed to the associations observed. We identified 567 DILI subjects with evidence of

autoimmune diseases; 135 of whom had a documented history of autoimmune/collagen vascular disease, and the remaining subjects were suspected to have an autoimmune disease because they had been treated with at least one drug commonly used in these conditions, usually in addition to the agent implicated as causing DILI (list of potential autoimmune treatments is presented in Table S11). When all 567 samples with known or suspected diagnosis of autoimmune disease were excluded from our cohort, the rs2476601 association remained highly significant with the same effect size (n=1245, OR =1.40 95%CI [1.23-1.60] P=$6.4*10^{-7}$, Table S12).

Assessment of correlation with known HLA risk alleles

We found an enrichment of rs2476601 among European DILI cases due to causal drugs known to have HLA alleles as the main genetic risk association (e.g., flucloxacillin, terbinafine, fenofibrate, minocycline, sertraline, AC) compared to the rest of the cases (OR =1.52 vs OR = 1.38, Table S13). Among these drugs associated with HLA risk alleles, AC was the major causal drug (444 cases) and showed the strongest association with rs2476601 (OR=1.62 95%CI [1.32-1.98] P = $4.0x10^{-6}$). AC-drug specific conditional analysis on HLA-A*02:01 and HLA-DRB1*15:01 confirmed that rs2476601 was an independently associated risk factor from the known HLA risk alleles (OR = 1.6 and P = $8.9x10^{-6}$, Table S5). Since the three markers were independent among each other, we looked for evidence of co-occurance of rs2476601 and the known HLA risk alleles. We therefore stratified AC cases and controls based on HLA allele carriage (Table S14). There was evidence that carriers of rs2476601 were enriched in AC DILI patients who carried one or both the HLA. In agreement with this finding, multi-marker analysis on the AC cohort confirmed that when rs2476601 co-occurred with either of the two HLA alleles, this consistently enhanced the association with DILI risk by almost two-fold compared to risk associated with the HLA alleles alone or in combination (Table 4). Joint carriage of the three markers was associated with a 13-fold higher DILI risk compared to the negative carriers. We had only 12 AC cases carrying only rs2476601 and neither of the two HLA risk alleles. This limited number did not allow us to capture the association of rs2476601 alone with AC DILI risk, but the 95% confidence interval reported in Table 4 included an OR of 1.5 . Moreover, when we compared the triple positive carrier against HLA-A*02:01 and HLA-DRB1*15:01 positive but rs2476601 negative group we found a significant 1.7 fold increase in the association with DILI risk (OR= 1.8 95%CI[1.24-2.60] ; P= 0.002). This confirmed that rs2476601 is

independently associated with AC DILI risk. Finally, we tested the presence of a SNP-HLA interaction effect for AC DILI. The analysis showed that there was an epistatic effect between rs2476601 and the presence of at least one of the HLA risk alleles (OR = 1.9; P = 0.05). In otherwords, the joint effect of one or both HLA risk alleles and rs2476601 was more than additive.

Multi-marker analysis on the terbinafine, flucloxacillin and flupirtine cohorts also supported rs2476601 as independently associated with DILI risk , showing consistent ORs in DILI due to flucloxacillin (OR=1.3), terbinafine (OR=3.4) and flupirtine (OR=2.5) in addition to the enhanced the DILI risk associated with the known HLA alleles (Table S15). We also examined whether subjects in the AC cohort who carried any combination of the three risk alleles differed in clinical phenotype (see Materials and Methods) from each other or from individuals not carrying any of the three alleles. No differences were apparent.

Discussion

Here, we report the results of the largest DILI GWAS to date based on 2048 DILI cases and 12,429 controls. Our analyses identified a robust association with a variant in *PTPN22*, a tyrosine phosphatase that has been linked to numerous autoimmune disorders[26]. Additionally, the association was not limited to a certain drug or pattern of injury, instead showing associations across the entire cohort. Moreover, we showed that the PTPN22 variant added to the DILI risk associated with known HLA alleles, increasing the association with AC DILI risk almost two-fold and appearing to have a similar effect on other DILI events with known HLA risk alleles. This finding provides new insights into DILI etiology and highlights the potential role of non-HLA variants in immune related genes as risk factors for DILI across a broad spectrum of causal drugs.

rs2476601 is associated with increased risk of type 1 diabetes mellitus, rheumatoid arthritis, systemic lupus erythematosus, vitiligo and Graves' disease, among others, but is also associated with decreased risk of Crohn's disease and Behçet disease[26]. This is the first confirmed genome-

wide association with DILI risk that lies outside the MHC locus and is the first variant that appears to generally predispose to DILI as opposed to DILI due to specific drugs. The replication of the association in separate Icelandic cases and controls where the PTPN22 variant is relatively common confirmed the association.

Our previous study also suggested that rs2476601 might be associated with risk for DILI, but the variant did not meet the criteria for genome-wide significance[7]. The present study confirmed that association, and with a larger sample size the strength of the association exceeded the required statistical threshold for a variant of this moderate effect size. It should be noted that the effect size seen for this variant with DILI was similar to those seen for other disease conditions where rs2476601 is associated with increased risk[27]. The frequency of rs2476601 varies greatly from population to population, being as high as 15% in Finns and as low as <0.01% in East Asians[28]. Here, we found that the frequency of the variant in DILI cases was higher in each population studied relative to the frequency in a matched control population, with the odds ratio remaining similar among different ethnic and racial populations. Despite this consistency of the odds ratio, the relatively small effect size of this variant means that larger sample sizes are needed to confirm the associations as real in all populations and with different agents responsible for causing DILI. The largest subset of the current study was Northern Europeans (n=1,107 cases and 5,090 controls), where a very strong signal for this variant was seen ($P=3.6x10^{-6}$, OR=1.41, Table S3). The other analyzed ethnicities, which had much smaller sample sizes, did not produce p-values below 0.05 despite their similar odds ratios for the effect of this variant.

Because patients with autoimmune disease may take more medications than others, there was a possibility that the association we observed with rs2476601 could actually be due to an increased prevalence of autoimmune diseases in DILI patients. To address this possibility, we identified cases with a recorded or suspected (based on concomitant medications) diagnosis of autoimmune diseases. Although rs2476601 in these patients had a slightly higher effect size, the association remained comparably strong even among patients not known or suspected to have underlying autoimmune conditions. Because history of autoimmune diseases was not systematically collected in all subjects, and because patients with autoimmune diseases may not be taking medication treatment for these conditions at the time of the DILI event, we cannot rule out the

possibility that some of our DILI patients had undiagnosed and untreated autoimmune conditions.

The association with rs2476601 was consistent across various phenotypes in our cohort, including injury patterns (cholestatic or mixed vs. hepatocellular), causal drugs, and strength of causality assessment. However, we did not find any association with other features, including DILI latency. We found that AC cases and other cases with the highest causality scores tended to have the highest frequencies for rs2476601. As DILI is a diagnosis of exclusion, it is unavoidable to have some uncertainty about the true cause of liver injury in certain patients, and so for real risk associations, we expect to see the strongest associations in those cases with the highest causality score. In our cohort, most AC cases were classified as having a high likelihood of DILI since the AC-DILI characteristic phenotypes have been well defined[29]. The higher allele frequency in AC cases likely reflects a higher proportion of patients who truly have DILI due to this drug. However, there may be an AC-specific genetic effect of rs2476601, as AC cases had a higher allele frequency than did other DILI cases even when restricting to the same causality probability categories.

Our analysis showed that rs2476601 appears to be associated with DILI risk regardless of which HLA alleles are associated with DILI risk. This is also the case with autoimmune diseases where rs2476601 is associated across diseases that are themselves associated with different HLA alleles[26]. This effect is consistent with the fact that PTPN22 controls events downstream from HLA presentation of neoantigen as summarized in Burn et al [30]. PTPN22 encodes lymphoid protein tyrosine phosphatase (Lyp) which is expressed exclusively in immune cells. Although the mechanisms whereby the rs2476601 variant reduces immune tolerance are not clear, Lyp is involved in T-cell receptor signaling, acting at several intermediate points in the signaling cascade. Lyp also appears to influence regulatory T-cell function. Considerable data now support the concept that DILI can result from a T-cell-mediated immune attack on the liver, presumably directed at HLA-presented neoantigens. This response may be initiated relatively frequently during treatment with drugs capable of causing DILI. However, clinically important liver injury does not occur in most of these patients because immune tolerance is activated, [31, 32].

We therefore suggest that rs2476601 predisposes to DILI by reducing immune tolerance. Our finding supports this hypothesis, since we found a significant genetic interaction effect among HLA risk alleles and rs2476601, detecting that the AC DILI risk associated with the joint carriage of HLA risk alleles and the variant is more that the sum of the risks associated with each single risk factor. Most of the currently reported associations for other diseases and the PTPN22 variant also show HLA associations, particularly HLA class II associations which is consistent with the strong association seen with AC-DILI but only a slightly increased frequency of the variant with flucloxacillin-DILI where the associated allele is class I. Moreover, the increased frequency of the rs2476601 variant in DILI cases where there was no apparent HLA association suggests the possibility of other HLA risk alleles yet to be discovered and/or a potential role for PTPN22 in non-T cell-mediated forms of DILI where other immune cells might be involved.

Although the association with rs2476601 was robust, its effect size was modest. The odds ratio averaged about 1.3 in the various ethnic groups identified here. However, in the 10% of the subjects who also carried the two known HLA risk alleles (HLA-A*02:01 and HLA-DRB1*15:01), the risk of DILI due to AC was increased over 13-fold. Given the rarity of serious liver injury due to AC, despite its widespread use, genotyping for risk management is probably not realistic. There may be instances when genotyping for this variant together with the identified HLA risk alleles could improve confidence in the causality assessment but this testing is not currently commercially available to our knowledge. It should also be noted that with drugs causing more frequent and severe liver injury, genotyping to identify 10% of a patient population at 13-fold increased risk of DILI might be reasonable.

In addition to the rs2476601 association, we also found evidence for a novel independent HLA risk factor in HLA-C*04:01. This association is interesting as the allele may be a risk factor across ethnicities and ADR clinical phenotypes. In fact, HLA-C*04:01 has previously been shown to be associated with nevirapine hypersensitivity in a Malawian population,[33] and in our cohort, the association was concordant across all three populations (Table S8) and across multiple drugs and herbal preparations (Table S7).

In conclusion, we have identified the variant rs2476601 in the PTPN22 gene as the first robust genetic risk association for DILI lying outside the MHC region. In addition, this association is the first to be associated across a broad range of implicated drugs and across different ethnic backgrounds. rs2476601 is therefore the first identified general risk association for DILI. The prior well-established association of this polymorphism with the risk of autoimmune diseases broadens the role of the immune system in DILI pathogenesis and may help inform future treatment and prevention efforts.

Acknowledgments

We are extremely grateful to Daniele Cusi (Hypergenes), Patrik K. Magnusson (Swedish Twin Registry) and Javier Martin (Spanish DNA bank) for provision of control data. The iDILIC team are very grateful to Arthur Holden (iSAEC) for his continuing support. Contributors to sample collection via DILIN, iDILIC, the Spanish DILI registry, EUDRAGENE, and DILIGEN are listed in the Appendix.

Figure Legend:

Figure 1. Manhattan plot displaying the association results of (A) the meta-analysis among the three major populations (Europeans, African Americans and Hispanics) (B) the meta-analysis after conditioning on the four main known HLA DILI risk alleles among the three major populations (Caucasians, African Americans and Hispanics). The results are reported for variants which had a consistent effect in Europeans and at least one of the two additional populations. SNPs shown in green have a significance level less than $5x10^{-6}$ and red have a significance level less than $5x10^{-8}$.

References

1. Gulmez SE, Larrey D, Pageaux GP, et al. Transplantation for acute liver failure in patients exposed to NSAIDs or paracetamol (acetaminophen): the multinational case-population SALT study. Drug Saf 2013;36:135-44.
2. Reuben A, Koch DG, Lee WM. Drug-induced acute liver failure: results of a U.S. multicenter, prospective study. Hepatology 2010;52:2065-76.
3. Bjornsson ES, Bergmann OM, Bjornsson HK, et al. Incidence, presentation, and outcomes in patients with drug-induced liver injury in the general population of Iceland. Gastroenterology 2013;144:1419-25, 1425 e1-3; quiz e19-20.
4. Sgro C, Clinard F, Ouazir K, et al. Incidence of drug-induced hepatic injuries: a French population-based study. Hepatology 2002;36:451-5.
5. Stevens JL, Baker TK. The future of drug safety testing: expanding the view and narrowing the focus. Drug Discov Today 2009;14:162-7.
6. Daly AK, Donaldson PT, Bhatnagar P, et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet 2009;41:816-9.
7. Lucena MI, Molokhia M, Shen Y, et al. Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. Gastroenterology 2011;141:338-47.
8. **Nicoletti P, Aithal GP,** Bjornsson ES, et al. Association of Liver Injury From Specific Drugs, or Groups of Drugs, With Polymorphisms in HLA and Other Genes in a Genome-Wide Association Study. Gastroenterology 2017;152:1078-1089.
9. Urban TJ, Nicoletti P, Chalasani N, et al. Minocycline hepatotoxicity: Clinical characterization and identification of HLA-B *35:02 as a risk factor. J Hepatol 2017;67:137-144.
10. Urban TJ, Goldstein DB, Watkins PB. Genetic basis of susceptibility to drug-induced liver injury: what have we learned and where do we go from here? Pharmacogenomics 2012;13:735-8.
11. Urban TJ, Shen Y, Stolz A, et al. Limited contribution of common genetic variants to risk for liver injury due to a variety of drugs. Pharmacogenet Genomics 2012;22:784-95.
12. Fontana RJ, Watkins PB, Bonkovsky HL, et al. Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. Drug Saf 2009;32:55-68.
13. Hayashi PH. Drug-Induced Liver Injury Network Causality Assessment: Criteria and Experience in the United States. Int J Mol Sci 2016;17:201.
14. Aithal GP, Watkins PB, Andrade RJ, et al. Case Definition and Phenotype Standardization in Drug-Induced Liver Injury. Clinical Pharmacology and Therapeutics 2011;89:806-815.
15. Danan G, Benichou C. Causality assessment of adverse reactions to drugs--I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. J Clin Epidemiol 1993;46:1323-30.
16. **Tryka KA, Hao L,** Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 2014;42:D975-9.
17. **Das S, Forer L, Schonherr S,** et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284-1287.
18. Zheng X, Shen J, Cox C, et al. HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics Journal 2014;14:192-200.
19. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
20. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010;11:288.
21. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.

22. **Gudbjartsson DF, Helgason H**, Gudjonsson SA, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015;47:435-44.
23. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 2008;40:1068-75.
24. Kong A, Steinthorsdottir V, Masson G, et al. Parental origin of sequence variants associated with complex diseases. Nature 2009;462:868-74.
25. Eggertsson HP, Jonsson H, Kristmundsdottir S, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet 2017;49:1654-1660.
26. **Stanford SM, Bottini N**. PTPN22: the archetypal non-HLA autoimmunity gene. Nat Rev Rheumatol 2014;10:602-11.
27. Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nat Med 2015;21:730-8.
28. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285-91.
29. deLemos AS, Ghabril M, Rockey DC, et al. Amoxicillin-Clavulanate-Induced Liver Injury. Dig Dis Sci 2016;61:2406-16.
30. Burn GL, Svensson L, Sanchez-Blanco C, et al. Why is PTPN22 a good candidate susceptibility gene for autoimmune disease? FEBS Lett 2011;585:3689-98.
31. Mosedale M, Watkins PB. Drug-induced liver injury: Advances in mechanistic understanding that will inform risk management. Clin Pharmacol Ther 2017;101:469-480.
32. Cho T, Uetrecht J. How Reactive Metabolites Induce an Immune Response That Sometimes Leads to an Idiosyncratic Drug Reaction. Chem Res Toxicol 2017;30:295-314.
33. Carr DF, Chaponda M, Jorgensen AL, et al. Association of human leukocyte antigen alleles and nevirapine hypersensitivity in a Malawian HIV-infected population. Clin Infect Dis 2013;56:1330-9.

Author names in bold designate shared co-first authorship

Tables

Table 1: Clinical characteristics of the samples in the three major DILI case population

CHARACTERISTICS	Europeans N=1806	Hispanics N=109	African Americans N=133
Clinical information			
Mean age, *years*	55	41	47
Female, %	56.5	56.8	76.6
Median alanine aminotransferase (range), *U/L*	774 (9- 15065)	843 (20-9108)	780 (47- 7001)
Median alkaline phosphatase (range) , *U/L*	290 (11- 6239)	266 (79 -2414)	265 (74- 2399)
Median latency (range), *days*	28 (1-7046)	58 (3- 2789)	51 (3-935)
Injury Type			
Cholestatic (%)	463 (26%)	12 (11%)	25 (19%)
Hepatocellular (%)	747 (41%)	73 (67%)	80 (60%)
Mixed (%)	465 (26%)	19(17%)	22 (16%)
Not available (%)	130 (7%)	5(5%)	6 (5%)
Total	**1806**	**109**	**133**

Table 2. Summary statistics for the univariate trans-ethnic meta-analysis of genome-wide associated variants

Marker	ALL			CAU			AA			HISP		
	OR	95% CI	P	OR	95%CI	P	OR	95% CI	P	OR	95%CI	P
rs2476601	1.44	1.28-1.62	$1.2x10^{-9}$	1.42	1.27-1.60	$9.6x10^{-10}$	1.94	0.73-5.18	0.19	1.91	0.94-3.89	0.07
rs72631546	1.84	1.47-2.31	$1.2x10^{-7}$	1.79	1.41-2.27	$1.92 x10^{-6}$	-	-	-	2.07	1.27-3.36	0.003
rs3129880	1.48	1.6-9.32	$1.2x10^{-20}$	1.56	1.44-1.69	$5.09 x10^{-26}$	0.92	0.67-1.27	0.62	0.97	0.68-1.4	0.89
HLA-B*57:01	2.19	1.84-2.61	$1.40 x10^{-18}$	2.24	1.94-2.6	$4.53 x10^{-27}$	1.64	0.36-7.4	0.52	0.72	0.17-3.02	0.65
HLA-DQB1*03:03	1.67	1.41-1.99	$7.48 x10^{-9}$	1.69	1.46-1.96	$1.27 x10^{-12}$	-	-	-	0.97	0.29-3.16	0.95
HLA-DRB1*15:01	1.37	1.22-1.53	$1.03 x10^{-7}$	1.40	1.27-1.54	$3.19 x10^{-11}$	0.85	0.43-1.69	0.65	1.16	0.68-1.99	0.58
HLA-C*06:02	1.40	1.23-1.59	$1.88 x10^{-7}$	1.45	1.3-1.62	$6.47 x10^{-11}$	0.92	0.58-1.48	0.74	1.38	0.77-2.48	0.28
HLA-DQB1*06:02	1.31	1.17-1.46	$1.47 x10^{-6}$	1.40	1.27-1.55	$3.15 x10^{-11}$	0.90	0.66-1.21	0.48	1.11	0.64-1.91	0.72

Odds ratios (OR), confidence intervals (95%CI) and p-values (P) are presented after correcting for population stratification with EIGENSTRAT axes within each major population.

Table 3. Association with rs2476601 for drugs with at least 3 case carriers in the European cohort and OR > 1. Drug results are ordered by p-value

DRUGS	# Cases	AF	OR	95%CI	P
Amoxacillin/Clavulanic Acid	444	0.13	**1.62**	**1.32-1.98**	**0.000004**
Terbinafine	15	0.20	**3.23**	1.29-8.1	**0.01**
Sulfamethoxazole/Trimethoprim	42	0.17	2.07	1.16-3.71	0.01
Methotrexate	9	0.22	3.34	1.09-10.16	0.03
Rofecoxib	6	0.25	4.08	1.05-15.82	0.04
Valproic Acid	16	0.18	2.43	0.99-5.95	0.05
Flupirtin	6	0.25	**4.43**	**0.98-20.05**	**0.05**
Fenofibrate	10	0.20	2.93	0.97-8.87	0.06
Erythromycin	11	0.18	2.89	0.95-8.78	0.06
Doxycycline	6	0.25	3.21	0.85-12.1	0.09
Pravastatin	6	0.25	3.21	0.83-12.47	0.09
Nimesulide	20	0.12	2.10	0.81-5.41	0.12
Cefuroxime	4	0.25	3.45	0.69-17.28	0.13
Ethinylestradiol/Levonorgestrel	7	0.21	2.53	0.71-9.05	0.15
Isoniazid	43	0.13	1.59	0.84-3.02	0.16
Celecoxib	9	0.17	2.37	0.69-8.19	0.17
Flucloxacillin	**195**	**0.11**	**1.24**	**0.90-1.71**	**0.18**
Nitrofurantoin	74	0.12	1.40	0.85-2.32	0.19
Piroxicam	5	0.20	2.85	0.60-13.68	0.19
Gabapentin	5	0.20	2.79	0.58-13.39	0.2
Cefazolin	21	0.14	1.59	0.67-3.8	0.3
Mercaptopurine	10	0.15	1.72	0.50-5.97	0.39
Imatinib	8	0.12	1.70	0.38-7.56	0.49
Ticlopidine	**5**	**0.10**	**2.01**	**0.24-16.73**	**0.52**
Atorvastatin	29	0.10	1.32	0.56-3.09	0.53
Minocycline	**32**	**0.11**	**1.29**	**0.58-2.86**	**0.53**
Interferon Beta-1a	4	0.12	1.90	0.21-16.79	0.57
Amiodarone	5	0.20	1.64	0.28-9.73	0.59
Diclofenac	66	0.10	1.17	0.67-2.04	0.59
Ibuprofen	15	0.10	1.36	0.41-4.52	0.62
Herbal and dietary products	58	0.10	1.19	0.65-2.17	0.58
Disulfiram	8	0.12	1.40	0.32-6.13	0.65
All Other Therapeutic Products	9	0.11	1.33	0.30-5.93	0.71
Nicotinic Acid	4	0.12	1.45	0.17-12.17	0.73
Lisinopril	5	0.10	1.45	0.17-12.17	0.74
Phenytoin	10	0.10	1.18	0.27-5.11	0.82

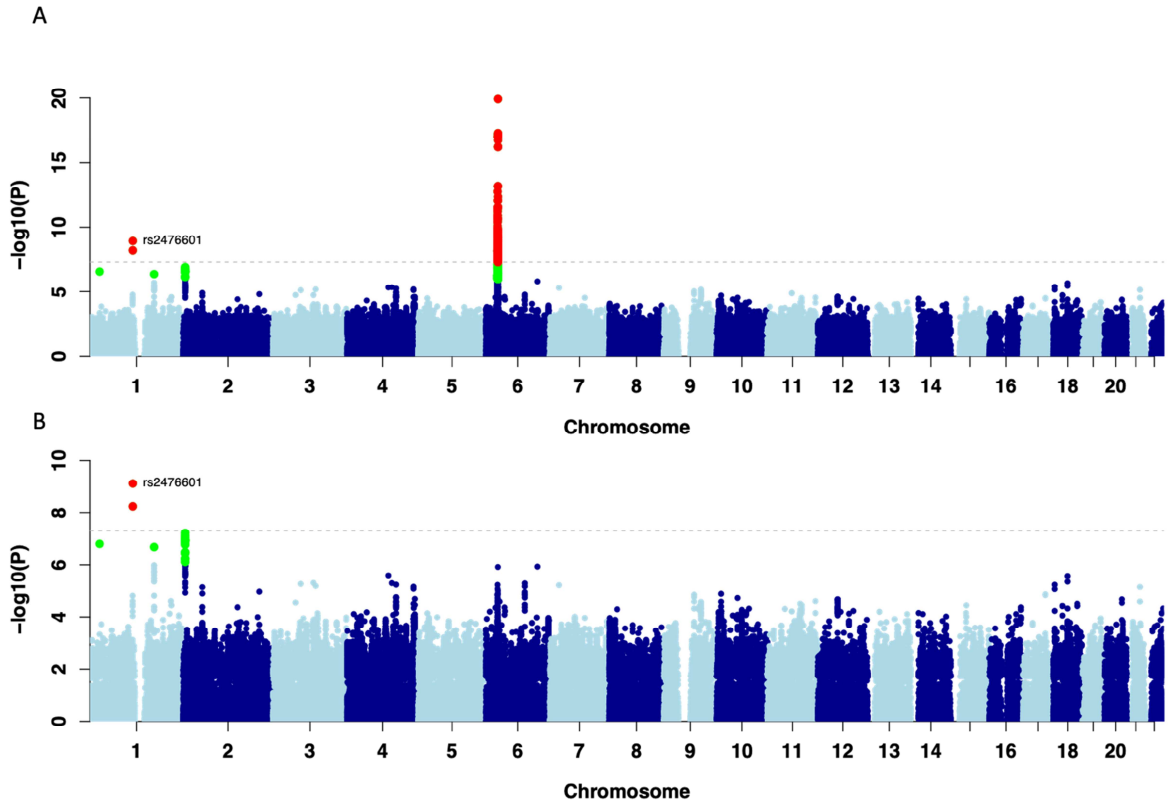
Rosuvastatin	4	0.12	1.27	0.14-11.29	0.83
Sertraline	6	0.08	1.17	0.15-9.27	0.88
Levofloxacin	17	0.09	1.05	0.32-3.44	0.94

Cases= number of cases for each drug; AF = Allele frequcncy; OR = Odd Ratio; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odds Ratio; P = logistic p-value. In bold drugs previously known to be associated with at least one HLA risk allele.

Table 4 Summary statistics of the multi marker analysis performed on the carriage of HLA-DRB1*15:01, HLA-A*02:01 and rs2476601 in the European Amoxicillin-Clavulanate-related DILI cohort.

	Cases		**Controls**				
Carriage group	**N**	**CF**	**N**	**CF**	**OR**	**95%CI**	**P**
+/+/+	49	0.11	187	0.02	13.80	9.18-20.71	$1.1*10^{-36}$
-/+/+	35	0.08	625	0.06	3.03	1.99-4.63	$2.3*10^{-7}$
+/-/+	12	0.03	206	0.02	3.28	1.74-6.19	$2.5*10^{-4}$
-/-/+	12	0.03	663	0.07	0.95	0.51-1.77	0.8
+/+/-	126	0.28	895	0.09	7.68	5.63-10.48	$8.8*10^{-38}$
-/+/-	101	0.23	3212	0.31	1.68	1.23-2.30	$1.1*10^{-3}$
+/-/-	41	0.09	1037	0.10	2.09	1.41-3.11	$2.7*10^{-4}$
-/-/-	68	0.15	3531	0.34	-	-	-

Odds ratios (OR), 95% confidence intervals (95%CI) and p-values (P) are presented after correcting for population stratification and considering the triple negative carriers as the reference group. The three letters in the first column (carriage group) reflects in order the HLA-DRB1*15:01 status, the HLA-A*02:01 status and the rs2476601 status. The risk alleles status is represented by "+ " = present or "–" = absent N = number of samples in the group; CF = carriage frequency.



Supplementary Material for Identification of a PTPN22 missense variant as a general genetic risk factor for drug-induced liver injury

Elizabeth T. Cirulli, Paola Nicoletti, Karen Abramson, Raul J. Andrade, Einar S. Bjornsson, Naga Chalasani, Robert J. Fontana, Pär Hallberg, Yi Ju Li, M Isabel Lucena, Nanye Long, Mariam Molokhia11, Matthew R. Nelson, Joseph A. Odin, Munir Pirmohamed, Thorunn Rafnar, Jose Serrano, Kari Stefansson, Andrew Stolz, Ann K. Daly, Guruprasad P. Aithal and Paul B. Watkins

page

Genome-wide association study Quality Control (QC) for each cohort

QC was conducted at both single marker and subject levels before performing the SNP imputation. Any marker that did not pass the following criteria was excluded from analysis: (i) genotype call rate in the batch of subjects greater than 95%, (ii) missing genotype rate greater than 5%, (ii) p-value for Hardy-Weinberg equilibrium greater than 10^{-7} in controls (if applicable). Any subject that did not pass the following criteria was excluded from analysis: (i) missing genotype rate < 0.05 among the SNPs that passed QC; (ii) not a sample duplicate or closely related based on estimated identity-by-descent (IBD) using PLINK v 1.07

Imputation

We used the Michigan Imputation Server[1] to impute missing genotypes separately for each ethnicity and for each subset genotyped by the same genotyping platform. We used the options of SHAPEIT for phasing, the Haplotype Reference Consortium as the reference for European ancestry samples and the 1000 Genomes Project as the reference for other ancestries[1-5]. Imputation methods are described in detail in the Supplementary Appendix. For HLA genotypes, four digit HLA alleles were inferred using HIBAG[6]. Sex chromosomes and mitochondria were not imputed. After imputation, the resulting calls were required to have $R^2 > 0.6$, maximum genotype posterior and genotype missingness < 0.05 in each group and the overall cohort. Genotypes were discretized based on the probability (PP) > 0.9. Variants were also removed if they were found to be heavily influenced by genotyping chip, as determined by a logistic regression p-value < 0.005 for a difference between two chip types within the case or within the control cohort.

Icelandic genetic analysis

The current association analysis was done on 113 DILI cases and 239,304 population controls using software developed at deCODE genetics[7] Genotypes of the Icelandic sample set were typed and then imputed as previously described [7,8, 9].The whole genomes of 15,220 Icelanders were sequenced, unveiling 40,780,213 single nucleotide polymorphisms (SNP) and short indels.

These variants were imputed into 151,677 Icelanders whose DNA had been genotyped with various Illumina SNP chips and phased using long-range phasing. Genealogical deduction of carrier status of 282,894 un-typed relatives of chip-typed individuals further increased the sample size for association analysis.. Logistic regression under an additive model was used to test for association between variants and disease, treating DILI as the response and expected genotype counts from imputation as covariates. Then those samples were used as reference for imputation of 155,250 Icelanders genotyped with chips. Using genealogic information, the sequence variants were also imputed into 282,894 relatives of the genotyped individuals [10]. HLA alleles were called for 28,075 Icelanders using whole genome sequence data and Graphtyper [11]. Association testing with multiple explanatory variables was performed using the *glm* function in R.

Table S1. Genotyping arrays used in each cohort.

Phenotype	Cohort	Ethnicity	n	Genotyping kit
Case	DILIN	European	296	Illumina 1MDuo
Case	DILIN	European	147	Infinium HumanCoreExome
Case	DILIN	European	389	Illumina MEGA
Case	DILIN	African American	32	Illumina 1MDuo
Case	DILIN	African American	101	Illumina MEGA
Case	DILIN	Hispanic	35	Illumina 1MDuo
Case	DILIN	Hispanic	74	Illumina MEGA
Case	iDILIC	European	361	Illumina 1MDuo
Case	iDILIC	European	508	Infinium HumanCoreExome
Case	iDILIC	European	105	Infinium OmniExpress
Control	Multiple sources	European	10397	Multiple
Control	MESA SHARe	African American	1314	Affymetrix Genome-Wide Human SNP Array 6.0
Control	MESA SHARe	Hispanic	718	Affymetrix Genome-Wide Human SNP Array 6.0

Table S2 Demographics, type of drugs leading to liver injury and the type of liver injury in the 113 genotyped Icelandic DILI patients.

Carattheristics	N of Cases (Frequency)
Demographics	
Age, median (Q1, Q3)	57 (41-71)
Females/males	51/62
Drugs	
Antimicrobial drugs	45/113 (40%)
Amoxicilin-clavulanate of Antimicrobials	30/45 (67%)
Type of injury	
Cholestatic/mixed type	72/113 (64%)
Hepatocellular type	41/113 (36%)

In the table median is reported with Q1= as 25th percentile and Q3 as the 75th percentile

Table S3. Allele frequencies of rs2476601 in different ancestry subsets of the analyzed samples and the gnomad database.

Ancestry group (case n / ctrl n)	OR	95% CI	P	AF Case	AF Controls	AF gnomad*
European ancestry	1.42	1.27-1.60	9.6x10-10	0.12	0.08	0.10
Northern Eur. (1,107 / 5,090)	1.41	1.22-1.63	3.6x10-6	0.13	0.09	-
Southern Eur. (209 / 2,518)	1.1	0.75-1.61	0.63	0.07	0.07	-
Swedish (146 / 1,076)	1.23	0.83-1.79	0.32	0.11	0.09	-
Jewish (87 / 1,076)	1.57	0.78-3.16	0.21	0.07	0.04	0.05
Other** (256 / 292)	2.36	1.45-3.85	0.001	0.11	0.05	-
Iceland (113/239,304)	1.48	1.09-1.99	0.01	0.13	0.08	-
African American (133/1,314)	1.94	0.73-5.18	0.19	0.02	0.01	0.01
Hispanic (109/718)	1.91	0.94-3.89	0.07	0.04	0.02	0.03

OR = Odds Ratio; 95% CI = 95% confidence intervals of the Odd Ratio; P = logistic p-value. AF = allele frequency. Note that the stated sample sizes do not add up to the total number included in this study because people who did not have their genotype successfully imputed for this variant were excluded. *The gnomad database[12] contains 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. **"Other" includes many different types of European ancestry as opposed to one particular cluster.

Table S4: Concordance rate between imputed and sequenced genotypes of rs2476601 by ethnicities (imputed vs typed)

A) Europeans

1:11437756 8_A	typed −9	0	1	2	Total
0	0	631	0	0	631
1	1	0	187	0	188
2	0	0	0	13	13
Total	1	631	187	13	832

B) African Americans

1:11437756 8_A	typed 0	1	Total
0	128	0	128
1	0	5	5
Total	128	5	133

C) Hispanics

1:11437756 8_A	typed 0	1	2	Total
0	95	0	0	95
1	0	8	0	8
2	0	0	1	1
Total	95	8	1	104

The genotypes are coded as 0=Homozygote minor 1=heterozygote and 2-homozigote major

Table S5: Summary statistics for known DILI risk factors in the multiple regression model within each ethnicity

Europeans	**Variant**	**OR**	**95%CI**	**P**
ALL cases	rs2476601	1.44	1.28-1.61	7.38E-10
	HLA-DRB1*15:01	1.47	1.33-1.63	3.98E-14
	HLA-A*02:01	1.28	1.18-1.38	9.41E-10
	HLA-B*57:01	2.42	2.08-2.80	4.53E-31
	HLA-A*33:01	2.21	1.60-3.04	1.38E-06
AC cases	rs2476601	1.61	1.31-1.99	8.91E-06
	HLA-DRB1*15:01	3.16	2.70-3.69	6.09E-47
	HLA-A*02:01	2.15	1.86-2.48	5.21E-26
African Americans	rs2476601	2.16	0.80-5.80	0.13
All Cases	HLA-DRB1*15:01	0.93	0.46-1.87	0.84
	HLA-A*02:01	0.58	0.36-0.91	0.02
	HLA-A*33:01	1.23	0.56-2.67	0.61
Hispanics	rs2476601	1.88	0.92-3.85	0.08
All cases	HLA-DRB1*15:01	1.20	0.70-2.05	0.52
	HLA-A*02:01	0.76	0.53-1.09	0.14
	HLA-A*33:01	0.93	0.41-2.12	0.87

AC cases = amoxicillin clavulanic acid cases; OR = Odds Ratio; 95% CI = 95% confidence intervals of the Odd Ratio; P = logistic p-value. AF = allele frequency. Odd ratio, confidence intervals and p-values are presented after correcting for population stratification with EIGENSTRAT axes and for other known HLA risk factors present in the populations.

Table S6: Summary statistics of rs2476601 and known HLA risk alleles in the Icelandic cohort

Analysis	Marker	OR	95%CI	P	AF
(a)Single marker[1]	rs2476601	1.52	(1.08-2.14)	0.016	
	HLA-DRB1*15:01	1.43	(1.06-1.92)	0.018	0.22
	HLA-A*33:01	1.33	(1.01-1.76)	0.045	0.30
	HLA-A*02:01	-		-	0.001
	HLA-B*57:01				0.04
(b) Conditioned to known HLA risk alleles	rs2476601	1.54	(1.10-2.17)	0.013	
	HLA-DRB1*15:01	1.43	(1.06-1.92)	0.018	0.22
	HLA-A*33:01	1.32	(1.00-1.75)	0.051	0.30
	HLA-A*02:01[2]	-		-	0.001
	HLA-B*57:01	1.22	(0.65-2.30)	0.541	0.04

OR = Odds Ratio; 95% CI = 95% confidence intervals of the Odd Ratio; P = logistic p-value. AF = allele frequency. Odd ratio, confidence intervals and p-values are presented after correcting for population stratification with EIGENSTRAT axes and with (a) and without (b) for other known HLA risk factors present in the population.

[1] Results are shown for markers with significant effect

[2]The frequency of HLA-A*02:01 is extremely low rendering the result in the joint model meaningless.

Table S7. The most associated HLA risk alleles after conditioning on the four known HLA DILI risk alleles.

HLA allele	OR	LCI	UCI	P	Effects
HLA-C*04:01	1.22	1.09	1.37	6.63E-04	+++
HLA-DRB1*04:01	0.78	0.67	0.90	9.30E-04	---

OR = Odds Ratio; ULC= 95% lower Confidence interval of the Odd Ratio; ULC= 95% upper Confidence Interval of the Odd Ratio; P = logistic p-value; Effects = effect in the three populations (Caucasians, African Americans and Hispanics) where + means a positive effect with OR > 1 and –means a negative effect with OR < 1.

Table S8. Statistics for HLA-C*04:01 in Europeans for drugs with at least OR > 1.2.

DRUGS	N	OR	LCI	UCI	P
Herbal and dietary preparations	58	2.24	1.40	3.59	0.0008
Methyldopa	5	8.09	2.28	28.67	0.001
Methotrexate	9	2.87	1.03	8	0.044
Nitrofurantoin	74	1.55	0.96	2.51	0.071
Lisinopril	5	3.78	0.89	16.02	0.071
Piroxicam	5	3.2	0.83	12.27	0.091
Atorvastatin	29	1.73	0.88	3.43	0.114
Amiodarone	5	3.06	0.56	16.82	0.198
Hydralazine	3	4.98	0.42	58.69	0.202
Sevoflurane	5	2.65	0.54	12.97	0.228
Gabapentin	5	2.41	0.5	11.62	0.274
Minocycline	32	1.47	0.71	3.02	0.297
Rosuvastatin	6	2.11	0.47	9.43	0.327
Omeprazole	5	2.02	0.41	9.81	0.385
Simvastatin	18	1.52	0.59	3.92	0.387
Pravastatin	6	1.99	0.42	9.51	0.387
Cefazolin	21	1.37	0.54	3.51	0.508
Sulfamethoxazole/Trimethoprim	42	1.25	0.64	2.44	0.509
isoniazid	43	1.21	0.62	2.34	0.575
Phenytoin	10	1.41	0.41	4.82	0.584
Azathioprine	37	1.22	0.59	2.54	0.598
Duloxetine	7	1.48	0.33	6.67	0.61
Moxifloxacin	8	1.24	0.28	5.44	0.772
Allopurinol	4	1.35	0.17	10.97	0.778
Interferon Beta-1a	4	1.31	0.14	11.77	0.812

OR = Odd Ratio, OR are from logistic regression including EIGENSTRAT axes as covariates.; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odds Ratio; P = logistic p-value

Table S9: List of agents in herbal and dietary supplements subgroup

Agents	N of cases
Unspecified Herbal	24
Other Combinations Of Nutrients	8
Herbal Nos W/Minerals Nos/Vitamins Nos	6
Camellia Sinensis	2
Garcinia Gummi-Gutta	2
Hydroxycut/Ephedra Free	2
Aloe Vera	1
Amino Acids Nos	1
Amino Acids Nos W/Capsicum Annuum Fruit/Chlor	1
Amino Acids Nos W/Herbal Nos/Minerals Nos/Vit	1
Carbohydrates Nos W/Creatine/Minerals Nos/Vit	1
Cimicifuga Racemosa	1
Ganoderma Lucidum	1
Ginkgo Biloba	1
Helianthus Tuberosus	1
Herbal Nos W/Minerals Nos	1
Herbal Nos W/Vitamins Nos	1
Herbals Nos W/Minerals Nos/Vitamins Nos	1
Trifolium Pratense	1
Uncaria Tomentosa	1
Grand Total	**58**

Table S10: Statistics for rs2476601 in Europeans stratifying the cases based on DILI phenotype.

Cohort	Number of cases	OR	LCI	UCI	P
HEPATOCELLULAR cohort	747	1.39	1.17	1.64	0.0001
CHOLESTATIC/MIXED cohort	927	1.50	1.30	1.74	6.50E-08

OR = Odd Ratio; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odds Ratio; P = logistic p-value

Table S11 List of drugs utilized in the treatment of the autoimmune diseases that served as a surrogate for suspicion of autoimmune disease in the DILI subjects.

Azathioprine / mercaptopurine, Cyclophosphamide, Cyclosporine, Hydroxychloroquine sulfate, Leflunomide, Methotrexate, Mycophenolate mofetil, Sulfasalazine/mesalazine, Apremilast, Tofacitinib, Tacrolimus, romiplostim and eltrombopag, levothyroxine, Propylthiouracil, methimazole, carbimazole, Neostigmine, etanercept, adalimumab, infliximab, certolizumab pegol, golimumab, Anakinra, abatacept, rituximab, and tocilizumab, canakinumab, Belimumab, prednisone, methylprednisolone, prednisolone

Table S12: Statistics for rs2476601 in Europeans stratifying the DILI cases based on whether cases had a predicted or reported diagnosis of autoimmune diseases previously associated with PTPN22 variant.

DRUG	#cases	OR	LCI	UCI	P	AF cases
Predicted autoimmune diseases based on drug history (reported in Table S8)	426	1.45	1.16	1.79	0.0008	0.12
Diagnosed with autoimmune disease	135	1.64	1.15	2.33	0.006	0.14
Both predicted and diagnosed autoimmune disease	561	1.49	1.24	1.80	2.95E-05	0.12
no evidence of autoimmune disease	1245	1.40	1.23	1.60	6.40E-07	0.12

#cases= number of cases, OR = Odds Ratio; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odds Ratio; PV = logistic p-value, AF cases =Allele Frequency in cases.

Note: the association analyses reported in the table has been performed using the same set of controls.

Table S13: Statistics for rs2476601 in Europeans stratifying the cases based on whether causal drugs were previously associated/not-associated with a HLA risk allele.

DRUGS	# Cases	OR	LCI	UCI	PV	AF cases
Drug associated with an HLA allele class I and class 2	719	1.52	1.29	1.80	8.53E-07	0.13
Drugs not-associatedd with an HLA allele class I and class 2	1279	1.38	1.19	1.59	1.09E-05	0.12

OR = Odds Ratio; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odds Ratio; PV = logistic p-value, AFcases =Allele Frequency in cases

Table S14: **Association of rs2476601 in European Amoxicillin-Clavulanic Acid cases stratifying based on the carriage of the two known HLA risk alleles.**

Groups	#cases	# Controls	OR	LCI	UCI	P	AF cases
ALL cases	444	10397	1.62	1.32	1.98	0.000004	0.14
++	178	1088	1.75	1.25	2.46	0.001	0.15
--	82	4208	1.11	0.64	1.94	0.7	0.09
+-	56	1250	1.5	0.82	2.73	0.2	0.12
-+	140	3851	1.69	1.17	2.44	0.005	0.13

Carriage of the HLA-DRB1*15:01 and HLA A*02:01 are expressed as "-" if absent and "+" if present following this order of HLA-RB1*15:01 as first digit and HLA A*02:01 as second digit. #cases = number of cases analyzed; #controls= number of controls analyzed; OR = Odds Ratio; ULC= 95% lower Confidence interval of the Odds Ratio; ULC= 95% upper Confidence Interval of the Odd Ratio; PV = logistic p-value; AF cases =Allele Frequency in cases

Table S15 Summary statistics of the multi marker analysis performed on the carriage of known HLA risk alleles and rs2476601 in the European DILI cohorts. 1) Terbinafine DILI cohort where the known risk allele is HLA-A*33:01, 2). Flucloxacillin DILI cohort where the known risk allele is HLA-B*57:01 and 3). Flupirtine DILI cohort where the known risk is the HLA-DRB1*16:01-DQB1*05:02 haplotype. The -/+ symbols in the first column reflect in order the known HLA risk allele status and the rs2476601 status.

1)Terbinafine	Cases		Controls				
	N	CF	N	CF	OR*	LCI	UCI
++	3	0.20	30	0.003	208.4	42.95	1011.30
-+	3	0.20	1,651	0.16	3.4	0.79	14.30
+-	4	0.27	184	0.02	50.1	11.97	210.13
--	5	0.33	8,491	0.82	-	-	-

2)Flucloxacillin	Cases		Controls				
	N	CF	N	CF	OR	LCI	UCI
++	34	0.17	109	0.01	77.96	44.70	135.97
-+	8	0.04	1,572	0.15	1.31	0.59	2.90
+-	125	0.64	665	0.07	55.91	36.36	85.99
--	27	0.13	8,050	0.77	-	-	-

3) Flupirtine	Cases		Controls				
	N	CF	N	CF	OR*	LCI	UCI
++	2	0.33	27	0.003	309.52	5.32	315.22
-+	1	0.17	1641	0.16	2.55	0.23	28.10
+-	1	0.17	260	0.03	16.07	1.45	177.80
--	2	0.33	8,357	0.80	-	-	-

Odds ratios (OR), 95% confidence intervals (95%CI) and p-values (P) are presented after correcting for population stratification and considering the double negative carriers (--) as the baseline group. N = number of samples in the group; CF = carriage frequency, *OR and CI are indicative of a trend since the number of cases from those drugs is very limited.

Figure S1: QQ plots for (a) the overall original analysis (b) conditional analysis.

(a) (b)

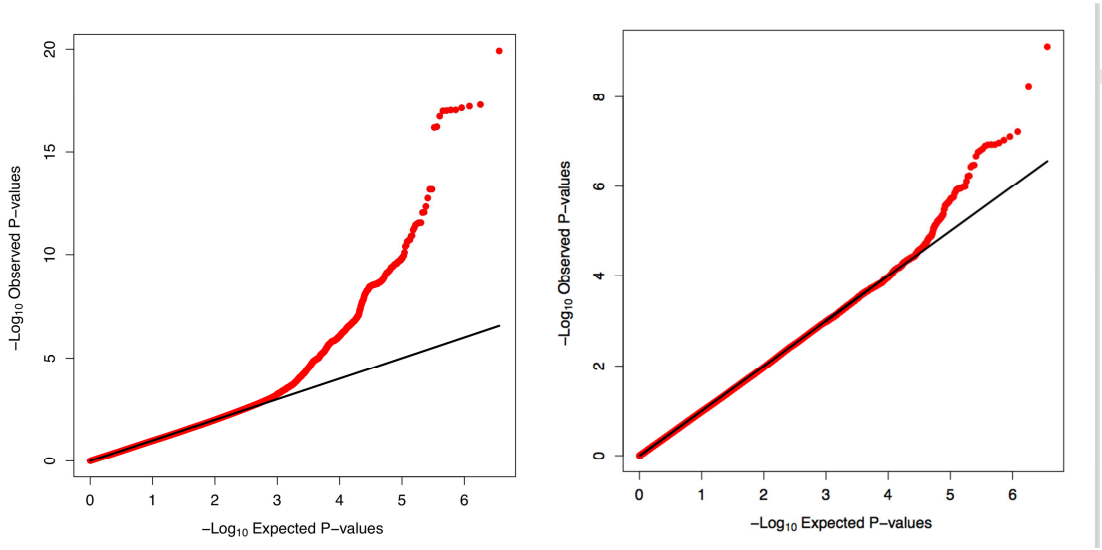
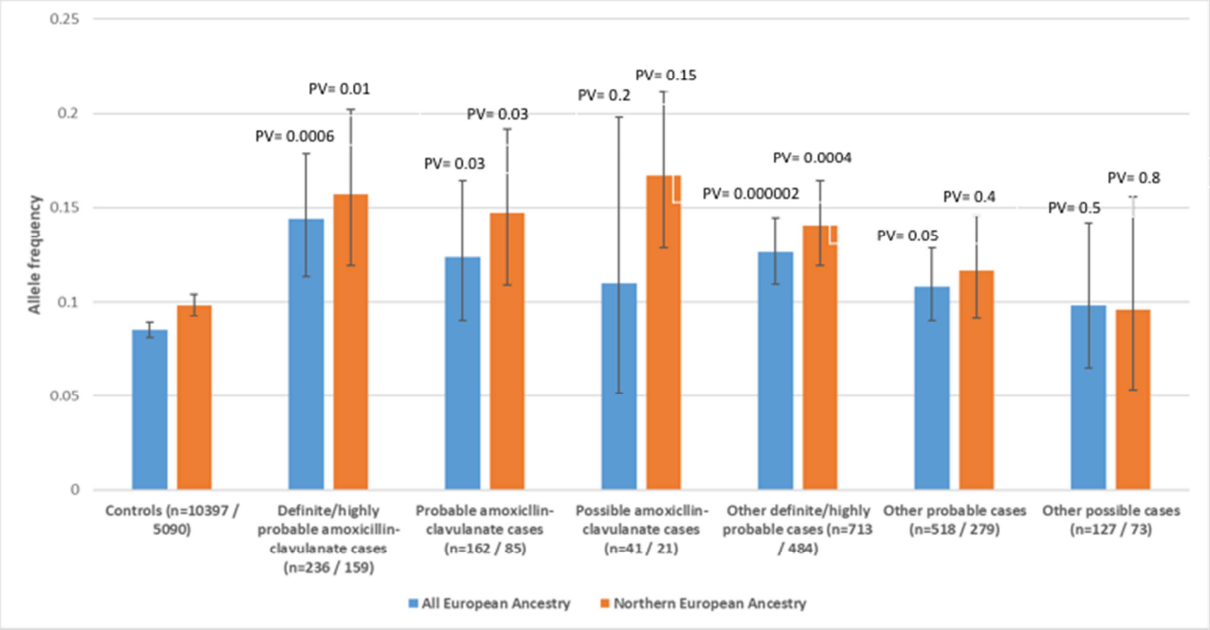


Figure S2. rs2476601 allele frequency across different subsets of our cohorts. Allele frequencies by causal drug and likelihood of DILI as described in methods. Error bars represent 95% confidence intervals.



The current stratification analysis based on causality score has been done dividing the cases by grouping DILIN “definite and highly likely” cases and iDILIC “highly probable” cases and grouping DILIN/iDILIC “probable” cases and grouping DILIN/iDILIC “possible” cases. PV= p-value

Collaborators and Contributors to case recruitment

iDILIC investigators (in alphabetical order)

Guruprasad P. Aithal, National Institute for Health Research (NIHR) Nottingham Digestive Diseases Biomedical Research Unit, Nottingham University Hospital NHS Trust and University of Nottingham, Nottingham, UK; Raul J. Andrade, IBIMA Hospital Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain and CIBERehd, Madrid, Spain; Fernando Bessone, Universidad Nacional de Rosario, Rosario, Argentina; Einar Bjornsson, Division of Gastroenterology and Hepatology, Department of Internal Medicine, The National University Hospital of Iceland, Reykjavik, Iceland; Ingolf Cascorbi, Institute for Experimental and Clinical Pharmacology, University Hospital Schleswig-Holstein, Kiel, Germany; Ann K. Daly, Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK; John F. Dillon, Ninewells Hospital and Medical School, Dundee, UK; Christopher P. Day, Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK; Par Hallberg, Uppsala University, Uppsala, Sweden; Nelia Hernández, Universidad de la Republica, Montevideo, Uruguay; Luisa Ibanez, Hospital Universitari Vall d'Hebron, Barcelona, Spain ; Gerd A. Kullak-Ublick, University of Zurich, Zurich, Switzerland; Tarja Laitinen, Helsinki University Central Hospital, Helsinki, Finland; Dominique Larrey, Hôpital Saint Eloi, Montpellier, France; M. Isabel Lucena, IBIMA Hospital Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain and CIBERehd, Madrid, Spain; Anke Maitland-van der Zee, AMC, Amsterdam, Netherlands; Jennifer H. Martin, University of Newcastle, Newcastle, NSW, Australia; Dick Menzies, MUHC and McGill University, Montreal Chest Institute, Montreal, Canada; Mariam Molokhia, King's College, London, UK; Munir Pirmohamed, Institute of Translational Medicine, University of Liverpool, Liverpool, UK; Shengying Qin, Shanghai Jiao Tong University, Shanghai, China; Mia Wadelius, Uppsala University, Uppsala, Sweden

DILIN investigators and coordinators can be found at http://www.dilin.org/publications/

References:

1. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284-1287.
2. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279-83.
3. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2011;9:179-81.
4. Delaneau O, Marchini J, Genomes Project C, et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 2014;5:3934.
5. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526:68-74.
6. Zheng X, Shen J, Cox C, et al. HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics Journal 2014;14:192-200.
7. Gudbjartsson DF, Helgason H, Gudjonsson SA, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015;47:435-44.

8. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 2008;40:1068-75.
9. Kong A, Steinthorsdottir V, Masson G, et al. Parental origin of sequence variants associated with complex diseases. Nature 2009;462:868-74.
10. Styrkarsdottir U, Thorleifsson G, Sulem P, et al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 2013;497:517-20.
11. Eggertsson HP, Jonsson H, Kristmundsdottir S, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet 2017;49:1654-1660.
12. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285-91.