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

Supplementary text ……2

Supplementary Tables ……..……..…. 4

Supplementary Figures …….……19

Contributors……. ……..... 17

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 ofrs2476601 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)

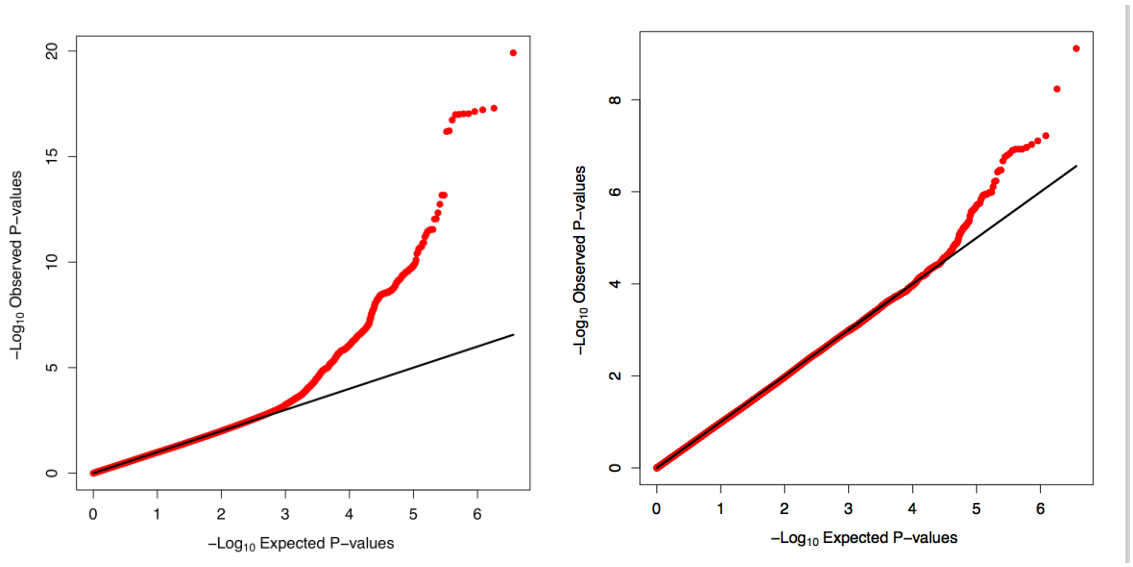
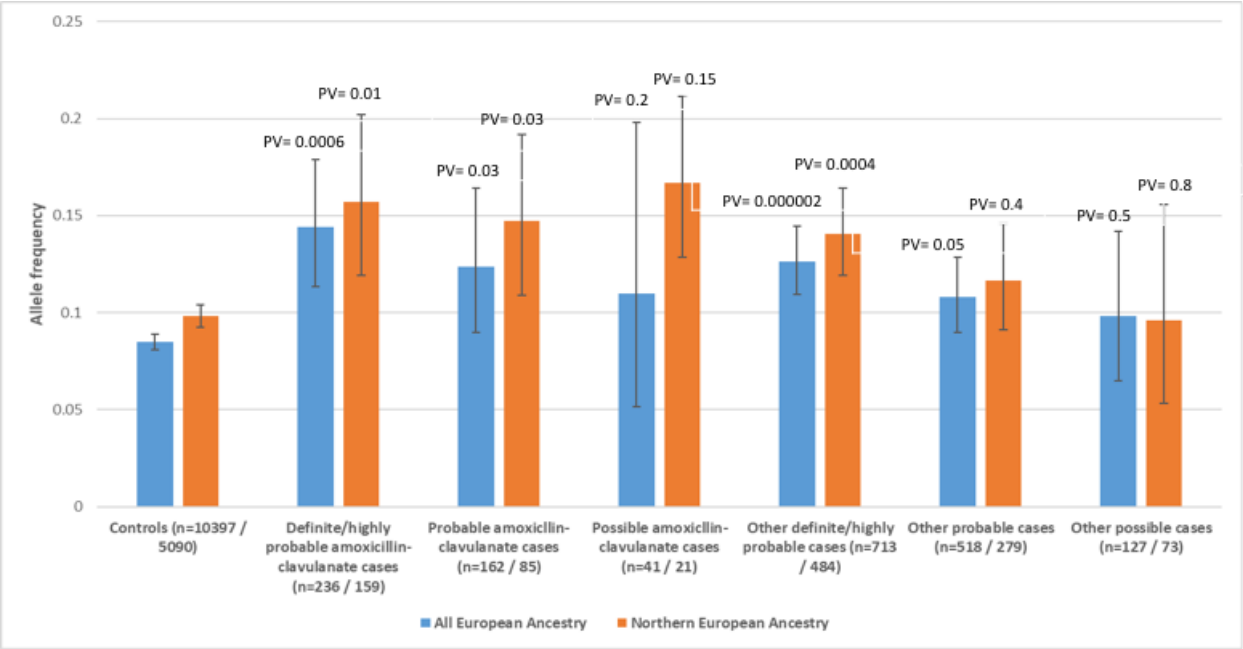


Figure S2. rs2476601 allele frequency across different subsets of our cohorts. Allele frequencies by causal drug and likelihood of DILI as described in methods. Error bars represent 95% confidence intervals.



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/

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.