Variability in Assigning Pathogenicity to Incidental Findings: Insights from *LDLR* Sequence Linked to the Electronic Health Record in 1013 Individuals

Maya S. Safarova MD PhD, Eric W. Klee PhD, Linnea M. Baudhuin PhD, Erin M. Winkler MS CGC, Michelle L. Kluge MS CGC, Suzette J. Bielinski PhD, Janet E. Olson PhD, Iftikhar J. Kullo MD

SUPPLEMENTARY MATERIAL AND METHODS ONLINE

STUDY PARTICIPANTS

A goal of the National Human Genome Research Institute's (NHGRI) electronic MEdical Records and GEnomics (eMERGE-III) Network is to assess penetrance and pathogenicity of variants identified by targeted sequencing of 109 disease-related genes. One of the aims of the Mayo eMERGE proposal is to determine which genetic variants are likely to contribute to familial hypercholesterolemia (FH), and therefore should be discussed with patients and families. This network was organized to combine DNA biorepositories with electronic health record (EHR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine on a national scale. Participants for this investigation were sampled from the eMERGE-PGx project in which a next-generation sequencing platform designed to assess sequence variation in 84 very important pharmacogenes (VIPs) was implemented in ~9,000 patients likely to be prescribed drugs of interest in a 1 to 3 years' time frame across several clinical sites in US (1). National Institutes of Health (NIH) Pharmacogenomics Research Network (PGRNseq) (2), a high-throughput next generation sequencing platform, captures the following VIPs ranked in the alphabetic order: *ABCA1, ABCB1, ABCB11, ABCC2, ABCG1, ABCG2, ACE, ADRB1, ADRB2, AHR, ALOX5, APOA1, ARID5B, BDNF, CACNA1C, CACNA1S, CACNB2, CES1, CES2, COMT, CRHR1, CYP1A2, CYP2A6, CYP2B6, CYP2C19, CYP2C9,*

CYP2D6, CYP2R1, CYP3A4, CYP3A5, DBH, DPYD, DRD1, DRD2, EGFR, ESR1, FKBP5, G6PD, GLCCl1, GRK4, GRK5, HLA-B, HLA-DQB3, HMGCR, HSD11B2, HTR1A, HTR2A, KCNH2, LDLR, MAOA, NAT2, NPPB, NPR1, NR3C1, NR3C2, NTRK2, PEAR1, POR, PTGIS, PTGS1, RYR1, RYR2, SCN5A, SLC15A2, SLC22A1, SLC22A2, SLC22A3, SLC22A6, SLC47A1, SLC47A2, SLC6A3, SLC6A4, SLCO1A2, SLCO1B1, SLCO1B3, SLCO2B1, TBXAS1, TCL1A, TPMT, UGT1A1, UGT1A4, VDR, VKORC1, ZNF423 (2). On the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published list of 56 clinically actionable genes there are six pharmacogenes from the PGRNseq platform, i.e. *CACNA1S, KCNH2, LDLR, RYR1, RYR2, SCN5A* (genes are listed in the alphabetic order). This report is focused on *LDLR* given our laboratory's experience in heritable lipid and lipoprotein disorders, in particular our interest in FH. The Mayo Clinic's contribution included 1013 residents of Olmsted County who were participants in the Mayo biobank (3). All subjects gave written informed consent, and the study protocol was approved by the Institutional Review Board at Mayo Clinic (Rochester, MN).

***LDLR* SEQUENCING**

The 101 base paired-end sequencing reads were processed using the Bioinformatics Genome Positioning System (GPS) (4) DNAseq secondary analysis workflow, which includes alignment, single nucleotide and small insertion/deletion variant calling, and annotation. FASTQ files were aligned to the hg19 reference genome using Novoalign (version V2.07.13), with realignment and recalibration performed using Genome Analysis Toolkit (GATK, version 1.6-7-g2be5704) Best Practices version 3. Germline variations were called with GATK's UnifiedGenotyper and filtered based on the variant quality score recalibration method.

Additional annotation provided included population frequency data from HapMap, 1K Genomes, ESP6500, and BGI200, dbSNP data. All positions were reported based on Genome Reference Consortium Human genome builds 37 and 38.

Quality control

A number of filtering criteria were performed to remove false-positive calls, and data quality and error rates were carefully evaluated. Following the initial variant call generation, we performed further filtering to identify high-confidence variants in the targeted sequencing data. The criteria included: (1) total reads; (2) library complexity (3) capture efficiency (4) coverage distribution: 80% at 20x was required for completion; (5) capture uniformity; (6) raw error rates; (7) Transition/Transversion ratio (Ti/Tv); (8) distribution of known and novel variants relative to dbSNP—typically < 7% (9) fingerprint concordance > 99%; (10) sample homozygosity and heterozygosity and (11) sample contamination validation. **Figure S1** illustrates the percentage of samples with per-base coverage at 5x, 30x, and 200x, as plotted along the length of *LDLR*. All samples had 30x coverage in all exons with nearly 25% of samples having 600x coverage in the exons. The per-exon average coverage for *LDLR* varied from 251 to 862 reads.

Following standard bioinformatics processing, in the first round of variant calling and selection we employed SnpEff (5) to annotate genetic variants. Our choice of SnpEff was determined by the need of ascertainment of a full exome call set, handling VCF files as input and continually-updated transcript databases, and support by Genome Analysis Toolkit (GATK). Variants were filtered based on SnpEff which attempts to classify the impact a nucleotide variant has on the underlying amino acid and peptide characteristics. Specifically, we filter for "impact=high" variants, which include variants causing a stop codon, disrupting a start or a stop

codon, impacting a splice acceptor site, a splice-donor site, or a rare codon triplet; “moderate impact” includes non-synonymous coding, codon insertion or deletion, 5' untranslated region (5′ UTR) deletion, deletion of 3'-UTR; “low impact” categorizes synonymous start/stop, coding, non-synonymous start/stop.

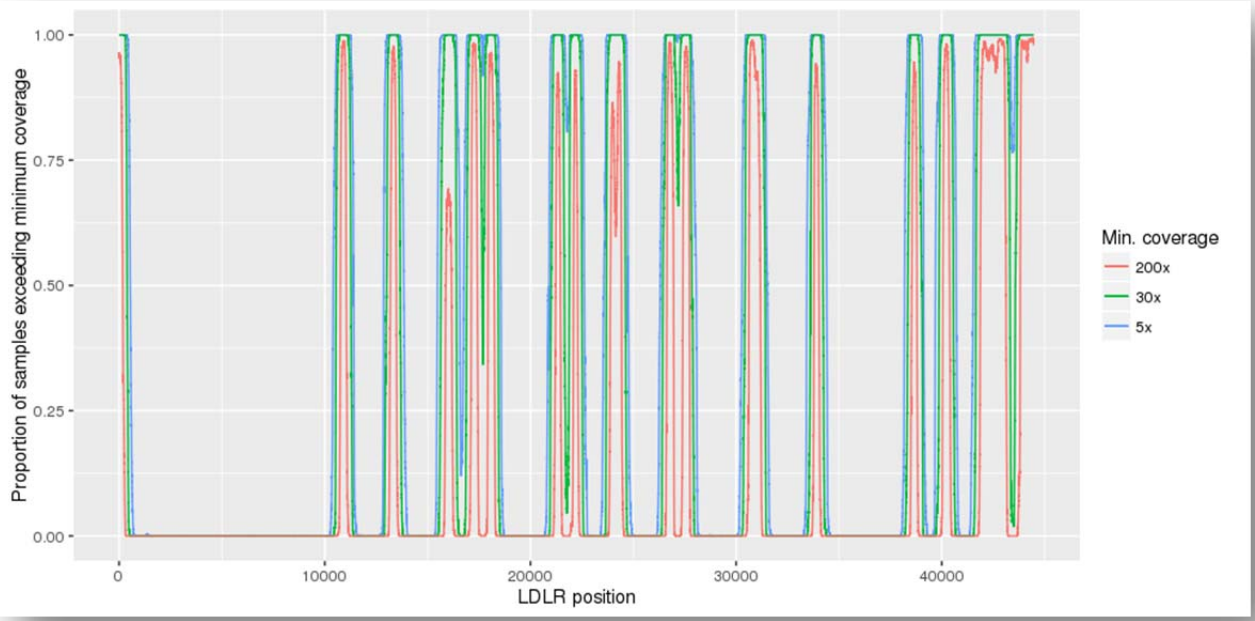


Figure S1. A plot of read depth per base for *LDLR* sequence data.

PHENOTYPING

Dutch Lipid Clinic Network criteria were used to identify patients at risk of FH-mutation carrier status (reviewed elsewhere (6)). This validated set of clinical criteria encompasses 5 diagnostic groups, i.e. family history and personal history of hypercholesterolemia or premature ASCVD, hallmarks of FH on physical examination, plasma LDL-C levels and molecular genetic testing. The highest applicable score is chosen per diagnostic group. A definite FH diagnosis can be made if the total score is >8 points, a probable FH is made when 6 to 8 points are scored, a

possible FH is referred to a range of 3 to 5 points, and an unlikely FH is made if the subject scores 0 to 2 points. The LDL-C levels were extracted from structured laboratory databases as the highest untreated value. None of the 124 individuals listed in **Table 1** were on a statin when LDL-C was ascertained. Each participant filled a questionnaire designed by Mayo biobank investigators to ascertain risk factor profile and family and personal medical history. We also reviewed "family history" and "patient provided information" sections of the EHR. A positive family history of early-onset ASCVD was found in 23 cases (22/124, 18%). **Table S1** (xcel spreadsheet labeled "Phenotyping") shows phenotypic diversity of individuals included in the current analysis. Details on each individual case with a positive family history of early-onset ASCVD are provided below.

Family history of premature ASCVD in 124 carriers of the selected 25 *LDLR* variants.

Variant	Family history of premature ASCVD
c.1691A>G (p.(Asn564Ser))	A 47-year-old male had a brother with a fatal myocardial infarction (MI) at the age of 38 years, his mother died of MI at age 59, his father had 4 strokes and MI starting around age 60 years
c.508G>A (p.(Asp170Asn))	A 60-year-old male, had a younger brother with MI at age 50 and his father underwent coronary bypass surgery at age 62 years
c.1171G>A (p.(Ala391Thr))	Out of 79 carriers of this variant, 10 had a positive family history. 1) A 56-year-old female had a mother with MI at age 57 years; 2) a 58 year-old female had a father with a fatal MI at

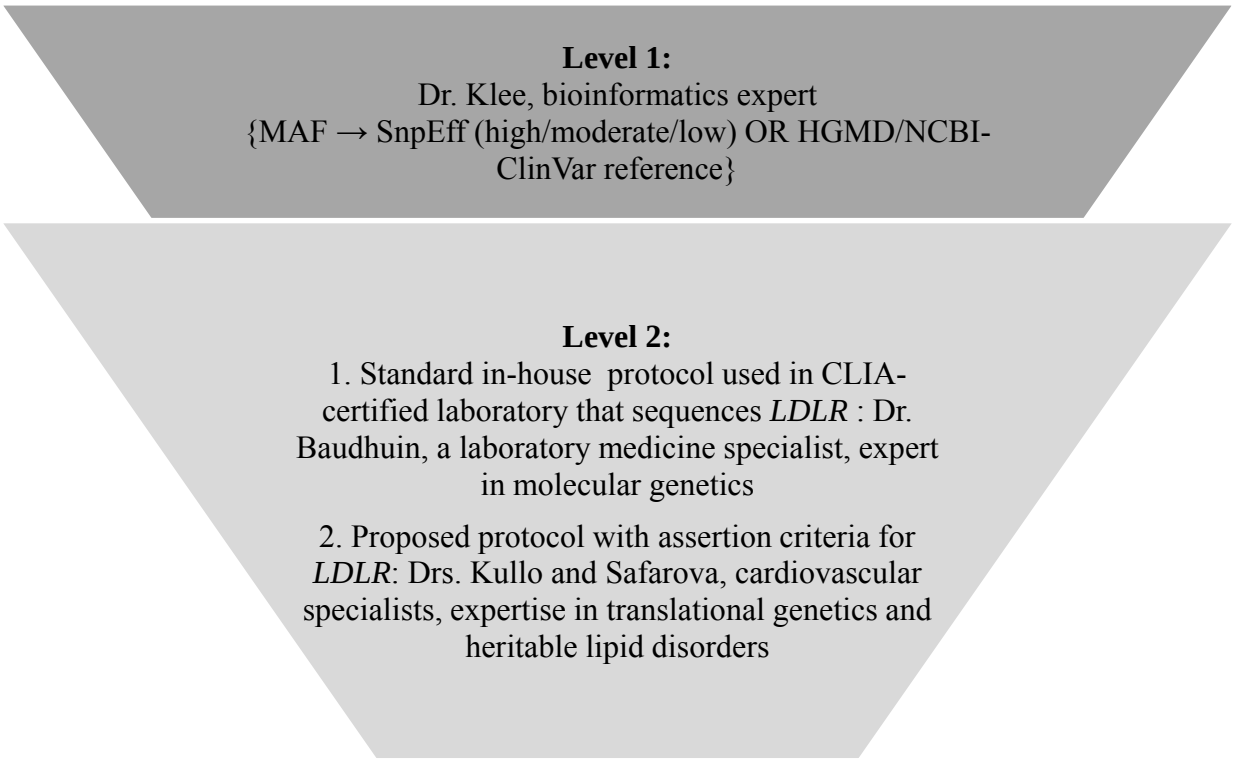
	age 53 years; 3) a 58-year-old female had a father diagnosed with CHD at early 50s followed by coronary artery bypass surgery (CABG) and a mother with 2 MIs before age of 65 years; 4) a 61-year-old female had a brother with a fatal MI at age 52 years; 5) a 52-year-old female had a mother with MI followed by CABG at the age of 55 years; 6) a 56-year-old female with a father who had a fatal MI at age of 48 years; 7) a 57-year-old male had a father with MI at age 54 years and a mother diagnosed with abdominal aortic aneurism at her early 60s; 8) a 64-year-old male had a mother with CHD at the late 50s; 9) a 60-year-old male with a father who experienced a fatal MI at age 54 years, 10) a 63-year-old male with a father with MI at the age of 46 years and stroke at the age of 48 years. Two individuals were also carriers of either rs138315511 or rs13306498, no family history of premature ASCVD was recorded in these cases.
c.1194C>T (p.(Ile398=))	A 64-year-old-male had a mother diagnosed with stroke at the age of 40 years.
c.1706-10G>A	There were 4 carriers of this variant with 2 having a positive family history, i.e. 1) a 60-year-old female with a sister diagnosed with CHD and treated with coronary stenting in her early 60s, 2) a 60-year-old male with both father and brother having MI prior to the age of 55 years.

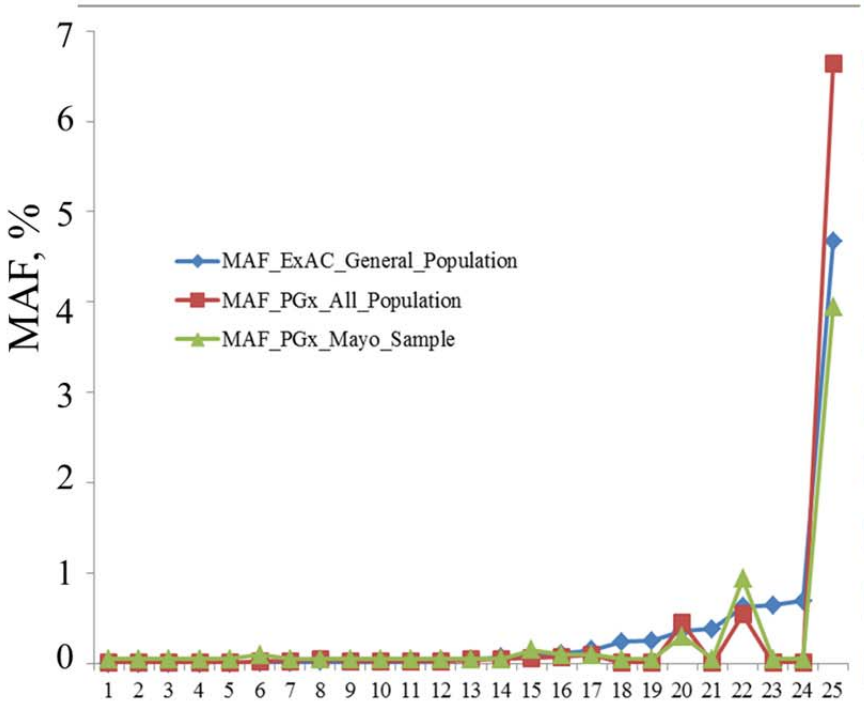
c.2177C>T (p.(Thr726Ile))	Out of 14 carriers of this variant, 7 subjects were reported to have a family history of premature of ASCVD. 1) a 59-year-old female had a mother with MI at age 44 years; 2) 49 female with a father with MI at age of 55 years; 3) 59 year-old female with a mother with MI at age of 50; 4) a 62-year-old male with a mother with MI at her early 60s; 5) a 61 year-old male with a father with fatal MI at age 49; 6) a 62-year-old male with a mother with MI in her early 60s; 7) a 61 year-old male with a brother documented to have a fatal MI at age 48.

On follow-up within the EHR surveillance timeline, a **personal history of premature ASCVD** was identified in three individuals: 1) a 53-year-old male, carrier of the missense variant c.2231G>A (p.(Arg744Gln)) experienced an early-onset transient ischemic attack, 2) a 67-year-old female-carrier of the c.1171G>A (p.(Ala391Thr)) variant had her first coronary event at the age of 62 years, and 3) a 59-year-old female-carrier of the c.1959T>C (p.(Val653Val)) variant was first diagnosed with CHD at age of 58 years during coronary calcium screening. The highest **LDL-C levels** were ascertained from the EHR and are presented in mmol/L in **Table S1**. In cases where the variant was present in more than one individual we calculated a median of all measurements to summarize this phenotype for the analysis. Further, 7 subjects were identified to carry 2 distinct *LDLR* variants. In 6 participants (c.1171G>A (p.(Ala391Thr)) was combined with i) missense c.2177C>T (p.(Thr726Ile)) or ii) synonymous c.1194C>T (p.(Ile398=), or iii) missense c.1085A>C1 (p.(Asp362Ala)). When their LDL-C levels were compared to those of single carriers of the c.1171G>A (p.(Ala391Thr)) variant no significant differences were

observed (P=0.3). The missense c.58G>A (p.(Gly20Arg)) variant occurred with c.2177C>T (p.(Thr726Ile)) in one subject leading to an untreated LDL-C level of 3.8 mmol/L measured at the age of 51 years. Compound-heterozygosity for c.1171G>A (p.(Ala391Thr)) with c.1085A>C2 (p.(Asp362Ala)) explained higher LDL-C levels (4.7 mmol/L) than in heterozygotes, whereas compound-heterozygous with c.1194C>T was associated with an LDL-C <100 mg/dL. None of the 25 *LDLR* variants described in the present report have been reported in the NHGRI-EBI catalog of published genome-wide association studies (7, 8), which is most likely explained by their low frequency.

VARIANT CALLING TEAM



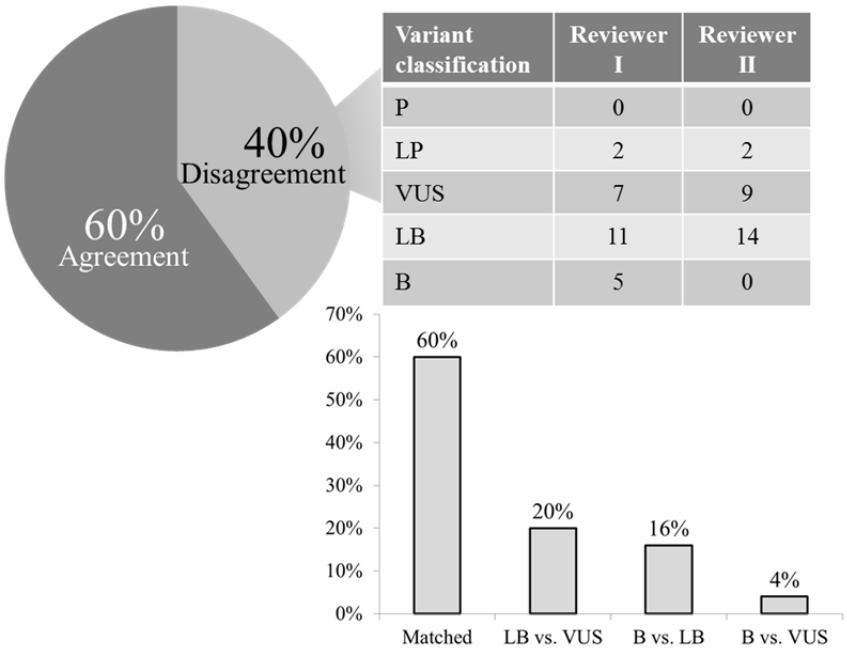


LDLR variants	PGx_Population	Mayo_Sample
rs137853960	9	1
rs137853962	4	1
rs139089530	1	1
rs143872778	4	1
rs143992984	5	1
rs144614838	1	1
rs146354103	5	1
rs148181903	1	1
rs148698650	8	1
rs150673992	9	1
rs200142970	4	1
rs758194385	1	1
rs368562025	1	1
rs116405216	1	1
rs12710260	4	1
rs72658861	4	1
rs72658867	4	1
rs137853963	18	2
rs138315511	5	2
rs147509697	13	2
rs13306498	11	3
rs17248882	10	4
rs5926	82	6
rs45508991	99	19
rs11669576	1213	80

Figure S2. Distribution of MAFs of the selected 25 *LDLR* variants. Blue line denotes data ascertained from the general population without race stratification; red line depicts total eMERGE-PGx cohort (n=9142) and green line represents a sample of subjects (n=1013) from Mayo Clinic. Minor allele frequency (MAF), number of alleles with the variant divided by the total number of

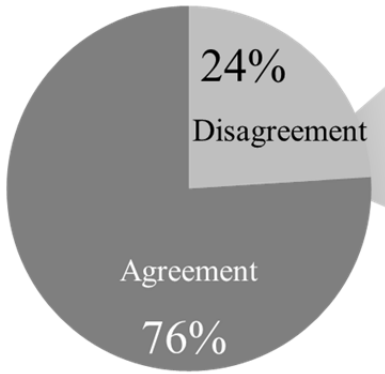
sequenced alleles, was calculated for each variant based on study data and the Exome Aggregation Consortium (ExAC) database (Cambridge, MA, URL: http://exac.broadinstitute.org, accessed 7/2015): http://exac.broadinstitute.org/transcript/ENST00000558013

A.



Variant classification	Reviewer I	Reviewer II
P	0	0
LP	2	2
VUS	7	9
LB	11	14
B	5	0

B.



Variant classification	Reviewer I	Reviewer II
P/LP	2	2
VUS	7	9
LB/B	16	14

Figure S3. Intra-reviewer concordance in classification of the 25 variants, using **A.** the five-level classification system, B. the three-level classification system by two independent expert reviewers: a laboratory medicine expert (reviewer I) and a cardiologist with an expertise in heritable lipid disorders (reviewer II). B=benign, LB=likely benign, VUS=variant of uncertain significance, LP=likely pathogenic, P=pathogenic.

The putative impact of variants was evaluated using an integrative summed score based on *in-silico* analysis (Table S2), with variants scored >1 considered to be likely benign, 0 – uncertain significance and <-1 – likely pathogenic.

Table S2 Bioinformatics prediction tools

PolyPhen-2		SIFT		MutationTaster		Mutationassesor		PROVEAN	
-1	Damaging/ probably damaging	-1	Disease causing	-1	Disease causing	-1	High/medium	-1	Deleterious
0	Possibly damaging	1	Polymorphism	1	Polymorphism	1	Neutral/low	1	Neutral
1	Benign								

***Footnote**: PolyPhen-2=Polymorphism Phenotyping v2, predicts possible impact of an amino acid substitution on the structure and function, annotates coding nonsynonymous polymorphisms based on the sequence, phylogenetic and structural information (9); SIFT=sorts intolerant from tolerant, predicts whether an amino acid substitution affects protein function based on the degree of conservation of amino acid residues in sequence alignments applied to nonsynonymous polymorphisms or (laboratory-induced) missense mutations (10); MutationAssessor predicts functional impact of a missense variant which assessed based on evolutionary conservation of the affected amino acid in protein homologs (11); MutationTaster integrates evolutionary conservation, splice-site changes, loss of protein features and changes that might affect the amount of mRNA followed by a naive Bayes classifier to predict the disease potential of 'silent' synonymous or intronic alterations, single amino acid alterations and complex changes in the amino acid sequence (12); PROVEAN=Protein Variation Effect Analyzer, predicts functional impact of single or multiple amino acid substitutions, insertions, deletions (13).*

Web Resources:

PolyPhen-2, http://genetics.bwh.harvard.edu/pph2, SIFT, http://sift.jcvi.org, *MutationTaster*, *https://omictools.com/mutationassessor-tool*, *Mutationassesor*, *http://www.mutationtaster.org/*, *PROVEAN*, *http://provean.jcvi.org/faq.php*

Table S3. Literature review was based on the analysis of the following original contributions:

rsID	cDNA position	Amino acid change	Description
rs758194385	c.1691A>G	p.(Asn564Ser)	Of 12 variants in *LDLR* sequence of 6503 participants, 9 were present in our annotated variant set (14), including this variant which was classified as pathogenic.
rs144614838	c.1432G>A	p.(Gly478Arg)	Of 12 variants in *LDLR* sequence of 6503 participants, 9 were present in our annotated variant set (14), including this variant which was classified as VUS. Reported as FH New York-2; FH Finn-9
rs137853962	c.1875C>T	p.(Asn625=)	Medeiros AM et al (15) reported this variant as likely benign.
rs137853960	c.148G>T	p.(Ala50Ser)	Reported as a VUS (14).
rs148698650	c.829G>A	p.(Glu277Lys)	Thormaehlen A et al (16): out of 13 missense variants in our study, we matched five with the list of 80 LDLR sequenced variants that underwent functional studies and were profiled based on the data from the Italian Study of Early-onset Myocardial Infarction (ATVB) cohort comprising 3235 individuals. The authors assigned pathogenicity based on the variant occurrence in patients with myocardial infarction as well as LDL-C ≥4.9 mmol/L, joint estimation of functional impact by four computational tools, including PolyPhen2 and SIFT, and FH-like disruptive-missense functional characterization. This variant was ascertained as benign (non-disruptive). Medeiros AM et al (15) reported this variant as benign.

rs368562025	c.1238C>T	p.(Thr413Met)	Reported as a VUS (14).
rs150673992	c.757C>T	p.(Arg253Trp)	Reported as a likely benign (14) and benign by Thormaehlen A et al (16).
rs200142970	c.2252G>A	p.(Arg751Gln)	Reported as a VUS (14).
rs148181903	c.1837G>A	p.(Val613Ile)	Reported as a VUS (14).
rs139089530	c.508G>A	p.(Asp170Asn)	Not reported
rs143992984	c.806G>A	p.(Gly269Asp)	Reported as a likely benign (14) and benign by Thormaehlen A et al (16).
rs146354103	c.507C>T	p.(Asn169Asn)	Not reported.
rs143872778	c.1836C>T	p.(Ala612=)	Not reported.
rs116405216	c.941-4G>A	-	Vandrovcova J et al (17) described thisvariant c.941-4G>A in FH.
rs138315511	c.1085A>C	p.(Asp362Ala)	This variant was associated with FH in the Danish and Dutch populations (18, 19). Reported as a VUS (14).
rs147509697	c.58G>A	p.(Gly20Arg)	This variant was found in the Polish individuals with FH (20). Reported as a likely benign (14) and benign by Thormaehlen A et al (16).
rs13306498	c.1194C>T	p.(Ile398=)	Reported to have a minimal effect on splicing (21).
rs137853963	c.2231G>A	p.(Arg744Gln)	Reported as benign by Thormaehlen A et al (16) as well as likely pathogenic as it has

			shown co-segregation with LDL-C levels and myocardial infarction in the German population (22).
rs11669576	c.1171G>A	p.(Ala391Thr)	Not reported.
rs12710260	c.1060+10G>C	-	Not reported.
rs17248882	c.1706-10G>A	-	Not reported.
rs45508991	c.2177C>T	p.(Thr726Ile)	Benign nature of these variants in the Spanish population was related by Tejedor M. (23, 24).
rs5926	c.1959T>C	p.(Val653Val)	
rs72658861	c.1061-8T>C	-	
rs72658867	c.2140+5G>A	-	Gretarsdottir S et al described this variant as expected to produce a truncated receptor, however at the same time to increase *LDLR* mRNA expression summarizing its protective role against CHD.

REFERENCES

1. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clin Pharmacol Ther* 2014; 96:482-9.
2. Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA and Scherer S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet Genomics* 2016.
3. Bielinski SJ, Olson JE, Pathak J, et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. *Mayo Clin Proc* 2014; 89:25-33.
4. Guo Y, Papachristoudis G, Altshuler RC, et al. Discovering homotypic binding events at high spatial resolution. *Bioinformatics* 2010; 26:3028-34.
5. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; 6:80-92.
6. Safarova MS and Kullo IJ. My approach to the patient with familial hypercholesterolemia. *Mayo Clin Proc* 2016; 91:770-86.
7. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; 42:D1001-6.
8. Burdett T, Hall PN, Hastings E, et al. The NHGRI-EBI Catalog of published genome-wide association studies. www.ebi.ac.uk/gwas.
9. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; 7:248-9.
10. Kumar P, Henikoff S and Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4:1073-81.
11. Reva B, Antipin Y and Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011; 39:e118.
12. Schwarz JM, Cooper DN, Schuelke M and Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014; 11:361-2.
13. Choi Y and Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015; 31:2745-7.

14. Amendola LM, Dorschner MO, Robertson PD, et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res* 2015; 25:305-15.
15. Medeiros AM, Alves AC and Bourbon M. Mutational analysis of a cohort with clinical diagnosis of familial hypercholesterolemia: considerations for genetic diagnosis improvement. *Genet Med* 2016; 18:316-24.
16. Thormaehlen AS, Schuberth C, Won HH, et al. Systematic cell-based phenotyping of missense alleles empowers rare variant association studies: a case for LDLR and myocardial infarction. *PLoS Genet* 2015; 11:e1004855.
17. Vandrovcova J, Thomas ER, Atanur SS, et al. The use of next-generation sequencing in clinical diagnosis of familial hypercholesterolemia. *Genet Med* 2013; 15:948-57.
18. Damgaard D, Larsen ML, Nissen PH, et al. The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. *Atherosclerosis* 2005; 180:155-60.
19. Fouchier SW, Dallinga-Thie GM, Meijers JC, et al. Mutations in STAP1 are associated with autosomal dominant hypercholesterolemia. *Circ Res* 2014; 115:552-5.
20. Chmara M, Wasag B, Zuk M, et al. Molecular characterization of Polish patients with familial hypercholesterolemia: novel and recurrent LDLR mutations. *J Appl Genet* 2010; 51:95-106.
21. Usifo E, Leigh SE, Whittall RA, et al. Low-density lipoprotein receptor gene familial hypercholesterolemia variant database: update and pathological assessment. *Ann Hum Genet* 2012; 76:387-401.
22. Braenne I, Kleinecke M, Reiz B, et al. Systematic analysis of variants related to familial hypercholesterolemia in families with premature myocardial infarction. *Eur J Hum Genet* 2016; 24:191-7.
23. Tejedor MT, Cenarro A, Tejedor D, et al. New contributions to the study of common double mutants in the human LDL receptor gene. *Naturwissenschaften* 2011; 98:943-9.
24. Tejedor MT, Cenarro A, Tejedor D, et al. Haplotype analyses, mechanism and evolution of common double mutants in the human LDL receptor gene. *Mol Genet Genomics* 2010; 283:565-74.