Atherosclerosis 281 (2019) 1–8

Contents lists available at ScienceDirect

Atherosclerosis

journal homepage: www.elsevier.com/locate/atherosclerosis

Systematic prediction of familial hypercholesterolemia caused by low-density lipoprotein receptor missense mutations

Jiayan Guo[b,1], Yan Gao[a,1], Xun Li[b], Ying He[b], Xin Zheng[a], Jianjun Bi[b], Libo Hou[a], Yinxi Sa[a], Mingqiang Zhang[b], Hong Yin[b,**], Lixin Jiang[a,*]

[a] *National Clinical Research Center of Cardiovascular Diseases, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China*

[b] *Amgen Biopharmaceutical Research & Development (Shanghai) Co., Ltd., Shanghai, China*

HIGHLIGHTS

- A total of 13,167 single amino acid missense substitutions in LDLR were modeled, in which 10,490 substitutions predicted FH pathogenicity with functional impact predictions.
- 52 out of all 54 reported experimentally-tested LDLR variants were consistent with our predictions.
- Three newly-identified FH-causing LDLR variants in patients and two novel benign variants validated our model predictions experimentally.
- This is the first to systematically predict pathogenic LDLR single amino acid missense substitutions with corresponding functional characteristics.

ARTICLE INFO

Keywords:
LDLR
Familial hypercholesterolemia
Function prediction
Missense mutation

ABSTRACT

Background and aims: Familial hypercholesterolemia (FH) is a an autosomal dominant disorder characterized by very high levels of low-density lipoprotein cholesterol (LDL-C). It is estimated that > 85% of all FH-causing mutations involve genetic variants in the LDL receptor (LDLR). To date, 795 single amino acid *LDLR* missense mutations have been reported in the Leiden Open Variation Database (LOVD). However, the functional impact of these variants on the LDLR pathway has received little attention and remains poorly understood. We aim to establish a systematic functional prediction model for *LDLR* single missense mutations.

Methods: Using a combined structural modeling and bioinformatics algorithm, we developed an *in silico* prediction model called "Structure-based Functional Impact Prediction for Mutation Identification" (SFIP-MutID) for FH with *LDLR* single missense mutations. We compared the pathogenicity and functional impact predictions of our model to those of other conventional tools with experimentally validated variants, as well as *in vitro* functional test results for patients with LDLR variants.

Results: Our SFIP-MutID model systematically predicted 13,167 potential *LDLR* single amino acid missense substitutions with biological effects. The functional impact of 52 out of 54 specific mutations with reported *in vitro* experimental data was predicted correctly. Further functional tests on LDLR variants from patients were also consistent with the prediction of our model.

Conclusions: Our LDLR structure-based computational model predicted the pathogenicity of *LDLR* missense mutations by linking genotypes with LDLR functional phenotypes. Our model complements other prediction tools for variant interpretation and facilitates the precision diagnosis and treatment of FH and atherosclerotic cardiovascular diseases.

[*] Corresponding author. National Clinical Research Center of Cardiovascular Diseases, State Key Laboratory of Cardiovascular Disease, FuWai Hospital, National Center for Cardiovascular Diseases, 167 Beilishi Rd., Beijing, 100037, China.

[**] Corresponding author. Amgen Biopharmaceutical Research & Development (Shanghai) Co., Ltd., 13F, Building 2, 4560 Jinke Road, Shanghai, 201210, China.

E-mail addresses: yinh@amgen.com (H. Yin), jiangl@fwoxford.org (L. Jiang).

[1] These authors contributed equally to this work as co-first authors.

https://doi.org/10.1016/j.atherosclerosis.2018.12.003
Received 18 May 2018; Received in revised form 26 November 2018; Accepted 4 December 2018
Available online 15 December 2018

1. Introduction

Familial hypercholesterolemia (FH) is a genetic disorder that can lead to premature atherosclerotic cardiovascular disease (ASCVD) and early death as a result of elevated plasma low-density lipoprotein cholesterol (LDL-C) levels [1]. FH often remains undiagnosed in asymptomatic patients until the development of ASCVD upon prolonged exposure to very high levels of plasma LDL-C [2]. Therefore, it is important to identify high-risk patients as early as possible. Consensus guidelines from international organizations (e.g., the European Atherosclerosis Society) recommend genetic cascade screening to identify pathogenic variants and begin early lipid-lowering treatment for the primary prevention of ASCVD [3,4].

The majority of FH variants are in the LDL receptor (*LDLR*) gene, in which missense variants account for approximately 46% [5]. Previous attempts to computationally predict the pathogenicity of LDLR variants have utilized different bioinformatics software, including MutationTaster [6], Polymorphism Phenotyping v2 (PolyPhen2 [7]), and Sorting Intolerant from Tolerant (SIFT [8]). However, predictions using these tools have been variable, with sometimes inconsistent results among programs or experimental findings; more importantly, these tools do not consider LDLR-specific biology, which is important for the precise diagnosis of FH. Thus, it is necessary to develop a precise computational model that can systematically predict LDLR variants and their associated functional impact.

After the cleavage of a 21-amino acid signal sequence, the mature human LDLR is an 839-amino acid type I transmembrane protein. To date, multiple LDLR domain structures have been solved, including LA1 to LA2 (Protein Data Bank [PDB] ID 1F5Y [9]), LA7 to EGF-C (PDB ID 3M0C), and LA2 to EGF-C at pH 5.3 (PDB ID 1N7D [10]). The functionally important extracellular part contains mainly three types of protein modules: 1) seven contiguous cysteine-rich repeats referred to as LDLR type A (LA) repeats (LA1-LA7), 2) three epidermal growth factor-like (EGF-like) repeats (EGF-A, EGF-B, and EGF-C), and 3) the YWTD (named for the conserved residues tyrosine, tryptophan, threonine, and aspartate) domain, which is between EGF-B and EGF-C. Structural and functional studies have demonstrated that the LA repeats are mainly responsible for binding to lipoproteins, while the EGF and YWTD domains are important for lipoprotein release at low pH and receptor recycling to the cell surface [11].

LDLR residues involved in disulfide bonds, calcium binding, and domain interfaces are crucial for the structural integrity and function of the receptor. Many of these residues have been identified as pathogenic sites for FH, including Cys329, which is located in a disulfide bond [12]; a calcium ion binding site near the carboxy-terminal end of the LA5 domain [13]; Asp362, Gln366, Arg574, Glu615, and Arg633, which are buried at the interface between the EGF-B and YWTD domain [10]; and His211 and His583, which are at the LA4-LA5 and YWTD domain interface [10].

Based on the LDLR structural resolution, we developed an improved *in silico* model called SFIP-MutID. It predicts the stratification of all single amino acid missense substitutions in the LDLR extracellular domain with biological characteristics and the likelihood of pathogenicity.

2. Materials and methods

2.1. Molecular modeling of LDLR under neutral and acidic pH

Structural models of the LDLR extracellular domain (ECD, amino acid coordinates: 22-714) were constructed to calculate the effect of single amino acid missense substitutions on protein stability. LDLR is known to have two states: an open conformation at a neutral pH for LDL binding and a closed conformation for LDL release at an endosomal pH in which the YWTD domain displaces bound LDL by interacting with LA4 and LA5 [11]. Accordingly, the LDLR structural models were built in these two conformations.

The neutral pH model was built using homology modeling in Discovery Studio [14] with the following templates: amino acids 22-104 from the nuclear magnetic resonance (NMR) structure of the LA1-LA2 domain (PDB ID 1F5Y [9]), amino acids 65-275 from the ECD structure (PDB ID 1N7D [10]), and amino acids 276-713 from the LDLR/proprotein convertase subtilisin/kexin type 9 (PCSK9) complex structure (PDB ID 3M0C). This model was refined with 30 nanoseconds of molecular dynamics simulation using the Desmond program (Schrödinger, New York, NY [15]) with the TIP3P explicit water model and 150 mM sodium and chloride ions added under the OPLS3 force field. The simulation was performed in the isothermal-isobaric (NPT) ensemble at a temperature of 300 K and a pressure of 1 bar. After simulation, the conformation with the lowest energy was selected as the final model. This conformation was protonated under a pH of 7.4 using the Prepare Protein protocol in Discovery Studio, and it was further minimized using the Generalized Born Molecular Volume solvent model with the Smart Minimizer algorithm and a maximum of 2000 steps.

Similar methods were used to build the LDLR ECD model under an acidic pH. Homology modeling was conducted with the following templates: amino acids 22-104 from the NMR structure of the LA1-LA2 domain (PDB ID 1F5Y) and amino acids 65-714 from the ECD structure solved at pH 5.3 (PDB ID 1N7D). The structure from PDB ID 1N7D was directly copied into the homology model. The model was protonated under a pH of 5.3 and further minimized as described previously.

2.2. Stratification and prediction of LDLR structure stability

The LDLR structural models were used to calculate the protein stability in Discovery Studio. Each site, except for 60 cysteine-forming disulfide bonds, was mutated to generate all possible single amino acid missense substitutions. For each mutant, the difference between the folding free energy of the wild-type and mutated structures was calculated. The larger value calculated from the two models, i.e., the neutral and acidic pH models, was used to predict the mutation effect on protein stability. The predicted effect of each substitution was categorized according to the default definitions in Discovery Studio: destabilizing if the mutation energy was greater than 0.5 kcal/mol, neutral if the mutation energy was −0.5 to 0.5 kcal/mol, and stabilizing if the mutation energy was less than −0.5 kcal/mol. Mutations in cysteines destroying disulfide bonds were also considered destabilizing due to the importance of disulfide bonds in protein folding and structure stability.

2.3. LDLR mutation architecture and annotation

All potential single amino acid missense substitutions in the ECD of the LDLR, including publicly reported and computationally simulated mutations, were mapped onto the LDLR sequence (amino acids 22-714) modeled in this study. A circular heatmap was drawn with ggplot2 [16] in polar coordinates of the x-axis (clock-wise, from 22 to 714). The reported frequencies of LDLR variants in large population-based studies were retrieved from the Exome Aggregation Consortium (ExAC) database [17]. Functional domains of the LDLR were based on the aforementioned ECD structure features.

2.4. Collection of LDLR missense variants for the in silico assessment of the accuracy of the model

Known missense mutations were identified using the UCL *LDLR* gene variant database (http://www.lovd.nl/LDLR [18]) with the pathogenicity update [19]. In this database, the predicted effects of missense variants on LDLR function were assigned according to the Association for Clinical Genetic Science (ACGS) guidelines (http://www.acgs.uk.com/media/774853/evaluation_and_reporting_of_sequence_variants_bpgs_june_2013_-_finalpdf.pdf), which define variants in classes 1 and 2 as likely and clearly non-pathogenic, class 3 as having unknown significance, and classes 4 and 5 as likely and clearly pathogenic. All variants reported in the database or that were computationally simulated are within the National Center for Biotechnology Information reference sequence NP_000518.1, GenInfo Identifier 4504975, which consists of 860 amino acids. For comparisons with our prediction model results, we considered ACGS classes 4 and 5 pathogenic and ACGS classes 1 and 2 non-pathogenic; class 3 variants were excluded from all comparisons. Pathogenicity predictions from MutationTaster, PolyPhen2, and SIFT were also extracted from LOVD.

2.5. Functional assessment of newly identified LDLR variants

2.5.1. Plasmid construction and cell culture

DNA sequences coding the human wild-type LDLR and mutant variants (p. Arg81His, p.Cys121Gly, p.Gly218Cys, p.Asp342Asn, and p.Arg595Trp) were cloned into the pIRESpuro3 vector (Clontech, CA, USA) with an HA tag sequence in the C-terminus by GenScript Co., Ltd. (Nanjing, China). HEK293 cells purchased from ThermoFisher Scientific were cultured in FreeStyle™ 293 Expression Medium (ThermoFisher Scientific, catalogue no. 12338018) at 37 °C and 5% CO_2. HEK293 cells showed no endogenous LDLR expression [20,21]. The cells were transfected with the plasmids, including empty pIRESpuro3 vectors, with Lipofectamine 3000 (Thermo Fisher Scientific, catalogue no. L3000015) according to the manufacturer's instructions. The transfected cells were cultured at 37 °C and 5% CO_2 for 48 h for fluorescence-activated cell sorting (FACS) analysis.

2.5.2. LDLR expression and ligand binding

Flow cytometry was used to detect cell surface LDLR expression, binding, and activity in the recycling pathway [20,22,23]. The fluorescence intensity was quantified using an LSRFortessa™ X20 (BD Bioscience, USA). All flow cytometry data were analysed using FlowJo software version 10.1. HEK293 cells (3×10^5/mL) transfected with wild-type and mutant plasmids were cultured in a 6-well plate for 48 h. The cells were harvested and incubated with a PE-mouse anti-human LDLR antibody (1:100, BD, catalogue no. 565653) at 4 °C for 30 min with light blocking. The cells were then washed and resuspended for flow cytometry measurement. To measure LDLR binding activity, cells were collected and incubated with 10 μg/mL Dil-LDL from human plasma (ThermoFisher, catalogue no. L3482) at 4 °C for 30 min with light blocking.

2.5.3. LDLR recycling activity

To measure LDLR activity in the recycling pathway, cells were incubated with 10 μg/mL LDL (Invitrogen, catalogue no. L3486) at 37 °C for 2 h. The cells were then washed and incubated with 0.5 mg/mL heparin at 4 °C for 1 h in the dark to release the LDL bound to the cell surface. The cells were washed three times and incubated with a PE-mouse anti-human LDLR antibody (1:100) at 4 °C for 30 min with light blocking. The cells were again washed and resuspended for flow cytometry measurement. The amount of LDLR in the recycling pathway, including internalization, reflects the cell surface LDLR after LDLR recycling, normalized to LDLR expression.

2.6. Statistical analysis

Statistical analysis was performed using GraphPad Prism 7. For the functional assessment of the five newly identified LDLR variants, multiple comparisons between the mutation groups and wild-type controls were tested using one-way analysis of variance followed by Dunnett's test. Differences were considered statistically significant if $p < 0.05$. The data are presented as the mean ± standard deviation.

3. Results

3.1. Structure-based functional impact prediction on single missense mutations

To stratify the pathogenicity of LDLR single missense mutations with functional impact, we assigned three types of scores to each mutation in the ECD (amino acids 22-714).

The first is a domain-based score. According to published data, the LA1 and LA2 domains are less “pathogenic” than the other LA domains (i.e., LA3 to LA7) and the EGF-A domain with regard to LDL binding [11]. In addition to LDL binding, LA 4 and LA5 are also involved in the interaction with the YWTD domain for LDL release. The EGF-AB and YWTD-EGF-C domain pairs are important for LDLR cell surface expression and recycling. The YWTD domain is indeed required to trigger the release of LDL at low pH, and quite a few FH mutations in this domain have been reported [11]. Based on these findings, a score of 0 was assigned to mutations in amino acids 22-106 (LA1-LA2), while a score of 2 was assigned to mutations in the LA4, LA5 and YWTD domains. The middle score of 1 was assigned to mutations in the other regions.

The second is the structural feature-based score. The LDLR ECD structure models under neutral and acidic pH are illustrated in Fig. 1. These models include the following categories of structural features: (1) cysteine residues forming disulfide bonds, (2) calcium ion binding residues, (3) intramolecular binding interface residues (acidic pH), (4) YWTD/EGF-B domain interface residues, (5) YWTD/EGF-C domain interface residues, (6) residues in the interior of the YWTD propeller, (7) conserved scaffolding residues within LA repeats (i.e., D57, D100, I122, D139, D178, I210, D227, I249, D266, I290, and D307) [24], (8) residues important for EGF-AB domain pair folding (i.e., G343 and R350) [25], (9) other surface-exposed residues, and (10) other non-surface-exposed residues (Supplementary Table 1). Changes in category (1), (2), (7), (8) and (10) residues are considered to be responsible for LDLR folding, and LA3-LA7 or EGF-A domain changes are responsible for LDL binding [11]. Category (3) residue changes can impair receptor recycling, and category (4), (5), and (6) changes affect LDLR folding and recycling. In combination with the domain assignment and published data [11], the possible disruptive impact (i.e., LDLR folding, LDL binding, and/or receptor recycling defects) of all residue changes was predicted for each domain (Supplementary Table 2). Residues with structural features, including disulfide bonds, calcium binding, and domain interfaces, are important for stability, and their mutations were given a score of 1. A score of 0 was assigned to mutations to residues without specific structural features (such as surface-exposed and non-surface-exposed) since there are more uncertainties in their predictions.

The third is the mutation energy-based score. “Destabilizing” mutations were given a score of 1, while a score of 0 was assigned to “Neutral” and “Stabilizing” mutations. The final score for each mutation was the sum of the three scores above. We considered mutants with final scores of 0 and 1 as unlikely and lowly pathogenic (i.e., benign) and those with final scores of 2, 3 and 4 as medium, highly and very highly pathogenic (i.e., pathogenic).

With this scoring algorithm, a total of 13,167 (693 sites with 19 possible amino acids) single amino acid missense substitutions were grouped into five pathogenic categories (Supplementary Table 3). The calculated data and predictions for the biological characteristics, such as folding defect, LDL binding defect, and/or recycling defect, of these single amino acid missense substitutions are presented in Supplementary Table 4. In total, 3381 of the 13,167 substitutions were predicted to be highly pathogenic: 3779 were predicted to be highly pathogenic, and 3330 were predicted to be medium pathogenic; the remaining 2677 substitutions were predicted to be lowly or unlikely pathogenic (2157 and 520, respectively) (Supplementary Table 3).

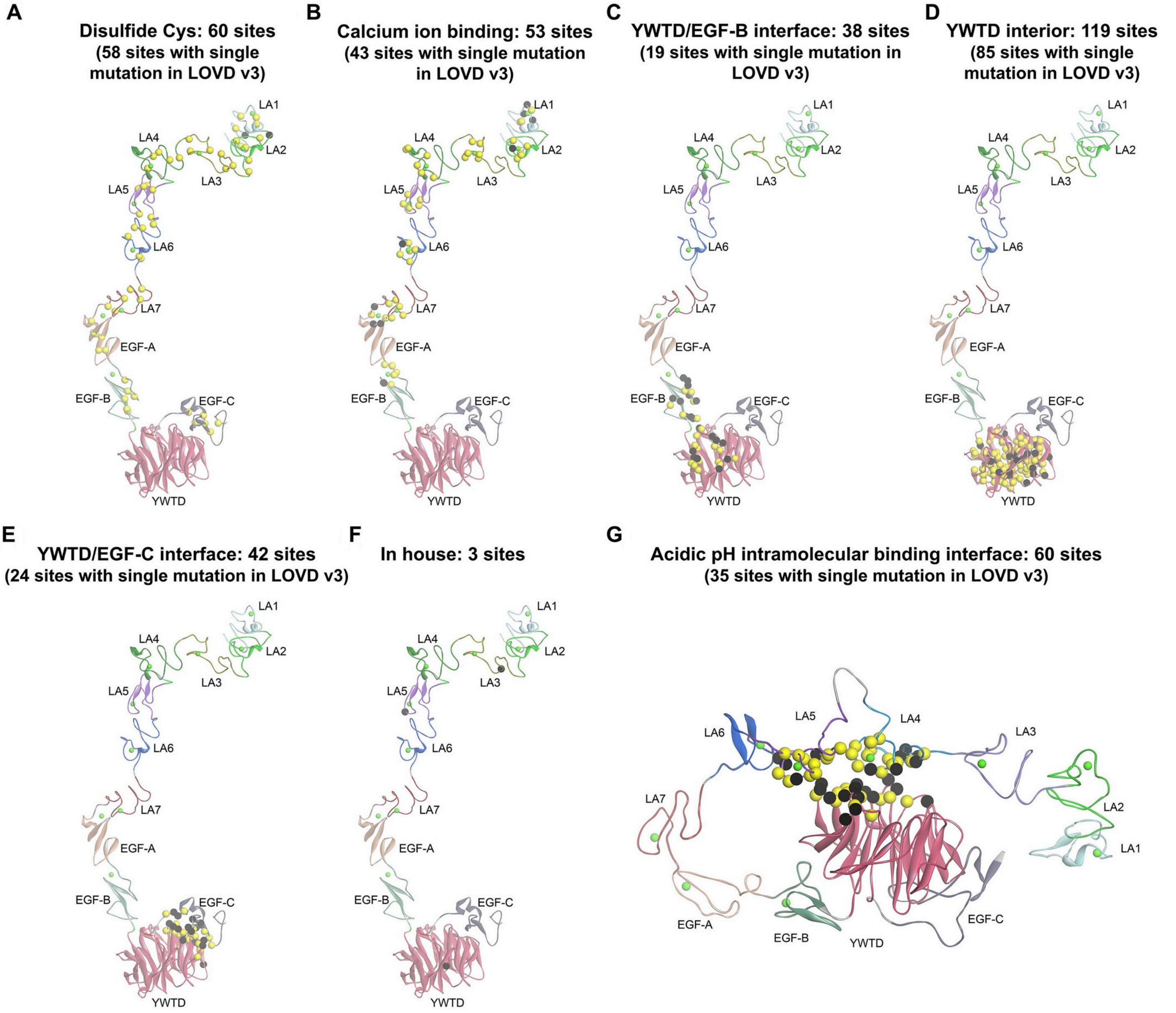
A
Disulfide Cys: 60 sites (58 sites with single mutation in LOVD v3)
B
Calcium ion binding: 53 sites (43 sites with single mutation in LOVD v3)
C
YWTD/EGF-B interface: 38 sites (19 sites with single mutation in LOVD v3)
D
YWTD interior: 119 sites (85 sites with single mutation in LOVD v3)
E
YWTD/EGF-C interface: 42 sites (24 sites with single mutation in LOVD v3)
F
In house: 3 sites
G
Acidic pH intramolecular binding interface: 60 sites (35 sites with single mutation in LOVD v3)
LA1
LA2
LA3
LA4
LA5
LA6
LA7
EGF-A
EGF-B
EGF-C
YWTD

Fig. 1. Mapping of FH mutation sites onto LDLR ECD (22-714) structure models.
(A–F) Mutations under a neutral pH; (G) mutations under an acidic pH. The LDLR is shown as a ribbon with labelled domains. The bound calcium ions are shown as green spheres, the FH mutation sites with single amino acid missense substitutions listed in LOVD are shown as yellow spheres, and other sites in the same structural categories are shown as black spheres. ECD: extracellular domain; EGF: epidermal growth factor; FH: familial hypercholesterolemia; LDLR: low-density lipid receptor; LOVD: Leiden Open Variation Database.

3.2. LDLR mutation architecture and distribution in the LDLR pathway

Our model construction and verification flow are shown in Supplementary Fig. 1. All 13,167 potential missense substitutions were mapped onto the *LDLR* sequence (amino acids 22-714) diagram (Fig. 2). In general, the LDL-C binding domains LA4, LA5, and YWTD were less tolerant of mutations and were enriched with missense pathogenic substitutions. LA1 and LA2, on the other hand, were enriched with mutations that were unlikely to be pathogenic. The interaction of PCSK9 with the EGF-A domain was more tolerant to missense substitutions than the LDL-C binding domains and its adjacent EGF-B domains. The distributions of all potential pathogenic substitutions (from medium to very high) in the LDLR pathway are displayed in Fig. 3. LA1 and LA2 were involved mainly in protein folding. LA4 and LA5 play important roles in multiple steps in the LDLR pathway, which is consistent with the mapping tolerability findings shown in Fig. 2.

3.3. In silico verification of LDLR variants

A total of 107 LDLR single missense mutation variants that underwent functional studies were identified from both published reports and LOVD (entries with “Enzyme Activity”). However, studies performed in compound heterozygous cells are considered to be invalid since it is not possible to determine the degree to which each allele contributes to the overall result [5]. In addition, not all 107 variants with *in vitro* experimental data were completely evaluated for all receptor functions. Therefore, we excluded unqualified variants and manually curated 54 variants (51 pathogenic and 3 benign missense mutations) for further analysis, as shown in Supplementary Table 5. Our model predictions were consistent with 52 out of 54 experimentally tested variants (Table 1). Comparisons with other prediction tools illustrate that our model precisely predicted pathogenicity (96.3%) with additional functional impact information (Table 1 and Supplementary Table 5).

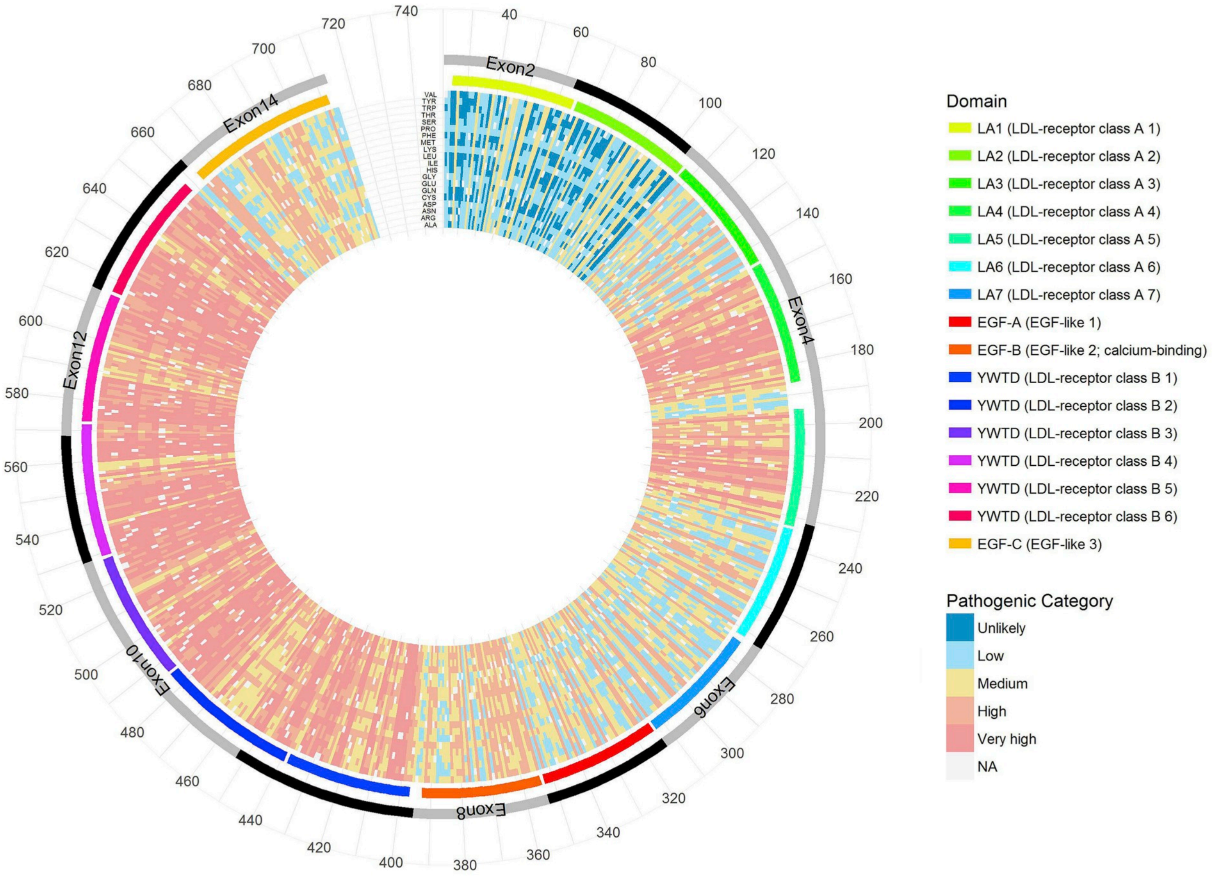

Fig. 2. Mutation architecture of the *LDLR* sequence.
All 13,167 potential missense substitutions, including publicly reported and computationally simulated substitutions, were mapped onto the *LDLR* sequence (amino acid [AA] coordinates: 22-714) modeled in this study. A circular heatmap was drawn using ggplot2 [16] in polar coordinates of the x-axis (clock-wise, from 22 to 714). For each amino acid, the mutation energies for substitution with the other 19 AAs are displayed. In the inner part of the plot (the heatmap part), each track represents one AA, each cell represents an AA site, and the colour of the cell reflects the mutation energy from the reference AA to the new AA on the track. In the outer parts of the plot (the dot part), the colour track represents one potential functional consequence of each substitution in the LDLR pathway (Supplementary Table 4); the grey/black track represents the corresponding exon and codon number. EGF: epidermal growth factor; LDLR: low-density lipoprotein receptor.

3.4. Experimental verification of LDLR variants

In a recent report [26], three FH patients with different LDLR variants (p.Cys121Gly, p.Gly218Cys, and p.Arg595Trp) that had not been functionally verified were diagnosed with probable and definite FH according to the Dutch Lipid Clinic Network criteria. To further verify the predicted functional impact of mutations according to our model, we conducted functional assays of the newly identified FH-causing LDLR variants [27,28]. In addition, two LDLR variants (p.Arg81His and p.Asp342Asn) considered to be benign by our prediction models and carried by two patients from the China PEACE cohort [26] (Dutch scores of 0 and 3, respectively) were selected for further functional validation.

A LDLR expression assay of the five newly identified FH-causing LDLR variants from patients revealed that the cell surface expression level of p.Cys121Gly was reduced to 32% compared with wild type. According to previous studies [5,29], results of less than 80% of the normal activity were considered a defect. The result suggested that the p.Cys121Gly variant may affect either LDLR synthesis or transportation to the cell surface (Fig. 4A). Variants carrying p.Gly218Cys and p.Arg595Trp showed normal expression (105.7% and 83.7%, respectively), while the binding capabilities were significantly diminished (46.8% and 59.0%, respectively) when normalized to their receptor expression (Fig. 4B). The LDLR recycling activities, including receptor internalization, were evaluated by treating cells with LDL and calculating the remaining cell surface LDLR. As shown in Fig. 4C, both p.Gly218Cys and p.Arg595Trp showed defects in the recycling pathway; this was particularly apparent for p.Gly218Cys. Variants carrying p.Arg81His and p.Asp342Asn showed normal activities in functional assays (Fig. 4).

4. Discussion

Our structure-based function prediction model provides an important improvement upon current *in silico* prediction tools to identify unknown FH-causing missense mutations with biological characteristics. However, the model should not be a “stand-alone” tool to replace any of the conventional analysis methods listed in Table 1. It should be used to complement additional supporting evidence used in the practises of variant interpretation (e.g., American College of Medical Genetics and Genomics (ACMG) guidelines). In addition, the level of confidence in predicting “high” and “very high” pathogenicity is greater than that for predictions of “unlikely” or “low” pathogenicity because “high” and “very high” predictions result from mutations with or within known structural features (e.g., disulfide bond, calcium binding, and domain interfaces); in contrast, the predictions of “unlikely” and “low” pathogenicity are more a result of the unknown or non-specific regions. Predictions in this regard have more uncertainties. Therefore, the strength of our model lies in predicting pathogenic mutations rather than predicting benign mutations.

The key features that differentiated from other tools enabled our prediction model to be a unique and complementary tool to interpret

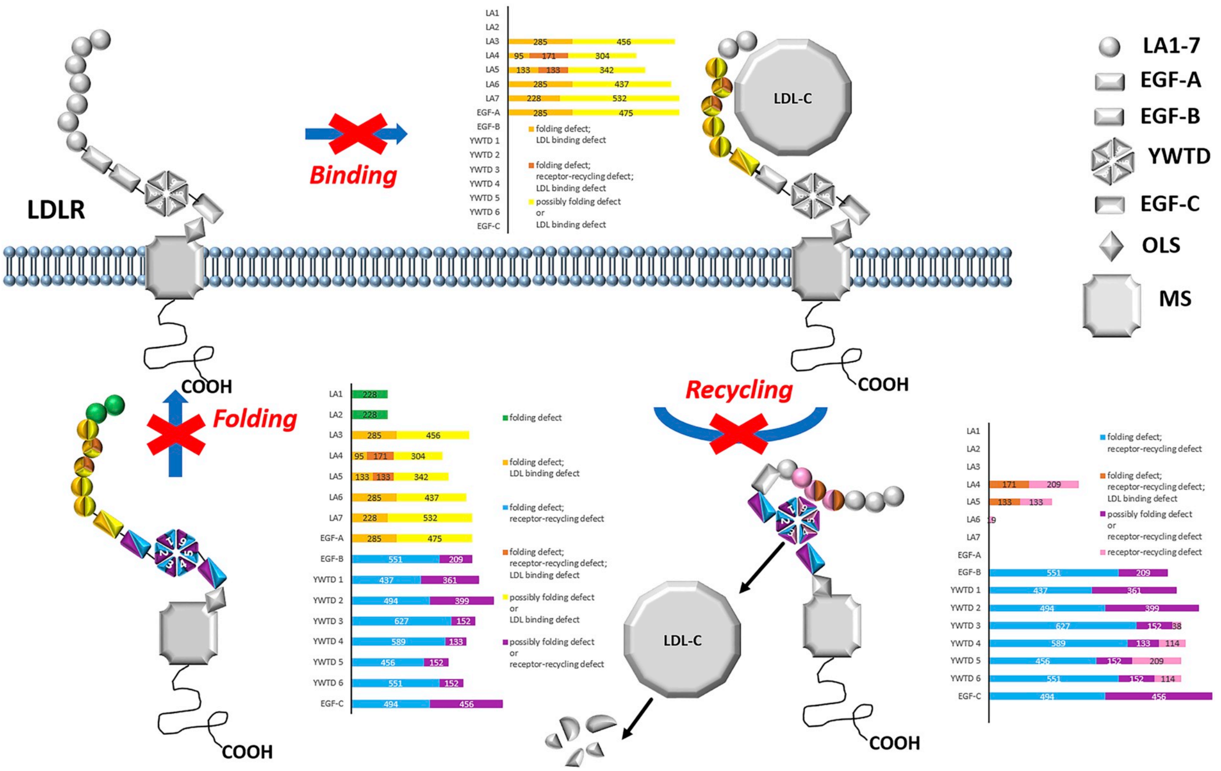

Fig. 3. Distribution of all potential pathogenic substitutions in the LDLR pathway.

The numbers of pathogenic substitutions are presented in the LDLR pathway and are characterized by their associated protein domain and functional categories in the LDLR pathway (Supplementary Table 4). For the three major steps in the LDLR pathway (i.e., LDLR folding, binding, and recycling), each functional domain of the LDLR is coloured based on the functional impact of the associated substitutions. EGF: epidermal growth factor; LA: low-density lipoprotein type A; LDL-C: low-density lipoprotein cholesterol; LDLR: low-density lipoprotein receptor; MS: membrane spanning; OLS: O-linked sugar.

Table 1

In silico pathogenicity prediction of experimentally validated missense mutations of LDLR variants.

Prediction tool	Predicted pathogenic (out of 51[a])	Predicted benign (out of 3[a])	Accuracy
ACGS from LOVD	51	3	100%
SFIP-MutID	49	3	96.3%
MutationTaster	49	1	92.6%
PolyPhen2	47	1	88.9%
SIFT	47	1	88.9%

ACGS: Association for Clinical Genetic Science; LDLR: low-density lipoprotein receptor; LOVD: Leiden Open Variation Database; PolyPhen2: Polymorphism Phenotyping v2; SIFT: Sorting Intolerant from Tolerant.

[a] Detailed information for the experimentally validated 51 pathogenic and 3 benign missense mutations is available in Supplementary Table 5.

variants with structural and biological annotations. For examples, MutationTaster uses DNA sequence conservation, mRNA stability, splice sites, and protein feature annotations for prediction; PolyPhen2 utilizes eight protein sequence features and three protein structure features for predictions; and SIFT focuses mainly on protein sequence conservation among homologs. Our prediction model considers not only structural features and domain/function annotation but also structure-based mutation energy; thus, it complements current strategies for predicting the pathogenicity of *LDLR* mutations and their functional impacts. Our predicted distribution of variants also verified the structural importance of certain LDLR domains, particularly LA4, LA5, and YWTD. Our results demonstrate a novel connection between these hotspot domains and their matching functional impact on the LDLR regulation pathway.

Our model had a high prediction accuracy when validated with experimentally validated LDLR variants. Only 2 of 51 pathogenic variants, Asp90Asn [30] and Gly137Ser [31], were not assessed correctly, both of which were predicted as benign (final score = 1) by our model. Asp90 is located in the LA2 domain, whose deletion has been reported to have little effect on LDL binding [11]; this information indicates that the mutation energy-based scoring needs further improvement. Gly137 is a surface-exposed residue without specific structural features, suggesting uncertainty for such residues.

In vitro functional studies of five newly identified LDLR variants from patients also validated the accuracy of our model. In our model, Cys121Gly and Gly218Cys were predicted to have high and very high pathogenic probabilities with functional impacts on binding and recycling activities, which is consistent with our functional results (Supplementary Table 6). Arg595Trp was predicted as a medium-level pathogenic variant with defects in folding or receptor recycling, which is also in accordance with our *in vitro* results. In contrast, Arg81His and Asp342Asn were predicted to have low pathogenic probability (benign), and our experimental results showed that the two variants presented with normal activities. Therefore, our computational model demonstrated the ability to provide precise predictions for not only the likelihood of pathogenicity but also the functional impact of the LDLR variants.

There is an increasing demand for the easy and precise genetic diagnosis of FH. This demand could be largely met by the precise computational systematic characterization of LDLR variants. For example, our model highlighted 3381 substitutions as having very high pathogenic probability with functional impact predictions. More importantly, the accurate functional classification of LDLR variants will largely facilitate precision diagnosis and appropriate treatment. With the newly developed PCSK9 inhibitors, which work through regulating the LDLR, the receptor expression levels of a particular FH patient may affect the drug response and therapeutic efficacy. For instance, *LDLR*

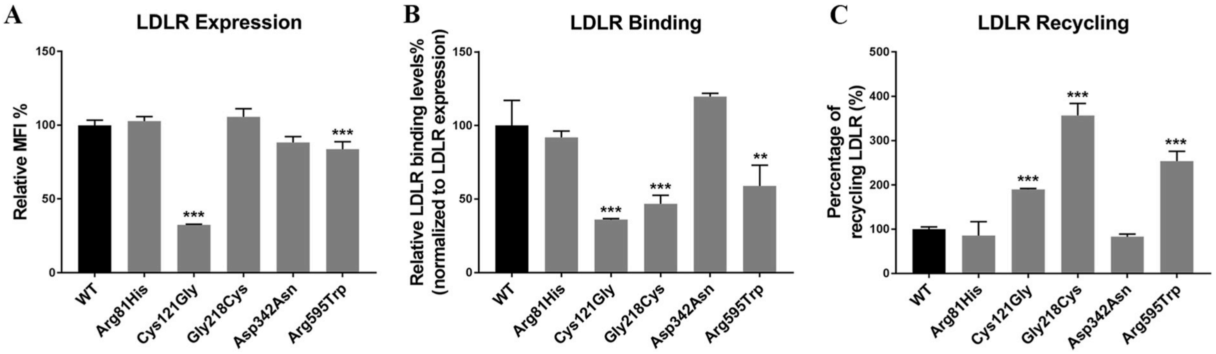

Fig. 4. Functional activities of LDLR variants from FH patients.
(A) LDLR on the cell surface was detected with a PE-mouse anti-human LDLR antibody by FACS. (B) LDLR binding activities were measured by treating transfected cells with Dil-labelled LDL from human plasma. (C) The amount of LDLR in the recycling pathway, including the internalization process, was evaluated by treating cells with LDL. All values represent the mean of triplicate measurements; the error bars represent ± standard deviation. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, compared with WT or healthy LDLR, one-way analysis of variance followed by Dunnett's test. FACS: fluorescent-activated cell sorting; FH: familial hypercholesterolemia; LDL: low-density lipoprotein; LDLR: low-density lipoprotein receptor; MFI: median fluorescence intensity.

polymorphisms (rs688, exon 12) can cause low LDLR expression and an impaired response to PCSK9 antibody treatment [32]. Our model is the first to systematically predict the functional impact of LDLR variants on the LDLR pathway and facilitate personalized diagnosis and treatment. Model development methods could also be applied for other causal genes in different genetic disorders.

However, there are still limitations to our model: 1) Only missense mutations, which account for 46% of the genotypes of LDLR variants [5], were predicted in our model. 2) Certain regions of the *LDLR* were not covered by our model (i.e., before codon 22 and after codon 714) due to the unresolved protein structure. 3) The mutation energy-based score still needs improvement. Therefore, combination usage with other prediction tools and experimental evidence will increase the accuracy of evaluating the pathogenicity and functional impacts of the LDLR variants.

Conflicts of interest

The authors declared they do not have anything to disclose regarding conflict of interest with respect to this manuscript.

Financial support

This work was supported by the Research Special Fund for the Public Welfare Industry of Health (201202025, 201502009) from the National Health and Family Planning Commission of China, the National Key Technology R&D Program (2015BAI12B01, 2015BAI12B02) of the Ministry of Science and Technology of China, the CAMS Innovation Fund for Medical Sciences (CIFMS, 2017-12M-2-002), and the 111 Project (B16005).

Author contributions

J.G., Y.G., M.Z., H.Y., and L.J. designed and conceived the study. Y.H. performed the variant analysis. X.L. constructed the computational model. J.G. and J.B. performed the functional experiments. Y.H., J.G., Y.G., X.L., X.Z., L.H., Y.S., M.Z., and H.Y. contributed to the collection and interpretation of the data and manuscript writing. All authors reviewed and approved the final manuscript.

Acknowledgements

We appreciate the great contributions of the China PEACE study group at the National Center for Cardiovascular Diseases and the cooperative hospitals across China with respect to the collection of clinical information and blood samples. We are grateful for the funding support provided by the Chinese government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atherosclerosis.2018.12.003.

References

[1] R.D. Turgeon, A.R. Barry, G.J. Pearson, Familial hypercholesterolemia: review of diagnosis, screening, and treatment, Can. Fam. Physician 62 (2016) 32–37.
[2] O. Najam, K.K. Ray, Familial hypercholesterolemia: a review of the natural history, diagnosis, and management, Cardiol. Ther. 4 (2015) 25–38.
[3] B.G. Nordestgaard, M.J. Chapman, S.E. Humphries, H.N. Ginsberg, L. Masana, O.S. Descamps, et al., Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society, Eur. Heart J. 34 (2013) 3478–3490a.
[4] M. Cuchel, E. Bruckert, H.N. Ginsberg, F.J. Raal, R.D. Santos, R.A. Hegele, et al., Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. A position paper from the Consensus Panel on Familial Hypercholesterolaemia of the European Atherosclerosis Society, Eur. Heart J. 35 (2014) 2146–2157.
[5] J.R. Chora, A.M. Medeiros, A.C. Alves, M. Bourbon, Analysis of publicly available LDLR, APOB, and PCSK9 variants associated with familial hypercholesterolemia: application of ACMG guidelines and implications for familial hypercholesterolemia diagnosis, Genet. Med. 20 (2018) 591–598.
[6] J.M. Schwarz, D.N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age, Nat. Methods 11 (2014) 361–362.
[7] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[8] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.
[9] N.D. Kurniawan, A.R. Atkins, S. Bieri, C.J. Brown, I.M. Brereton, P.A. Kroon, et al., NMR structure of a concatemer of the first and second ligand-binding modules of the human low-density lipoprotein receptor, Protein Sci. 9 (2000) 1282–1293.
[10] G. Rudenko, L. Henry, K. Henderson, K. Ichtchenko, M.S. Brown, J.L. Goldstein, et al., Structure of the LDL receptor extracellular domain at endosomal pH, Science 298 (2002) 2353–2358.
[11] H. Jeon, S.C. Blacklow, Structure and physiologic function of the low-density lipoprotein receptor, Annu. Rev. Biochem. 74 (2005) 535–562.
[12] L. Jiang, L.Y. Sun, Y.F. Dai, S.W. Yang, F. Zhang, L.Y. Wang, The distribution and characteristics of LDL receptor mutations in China: a systematic review, Sci. Rep. 5 (2015) 17272.
[13] D. Fass, S. Blacklow, P.S. Kim, J.M. Berger, Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module, Nature 388 (1997) 691–693.
[14] D.S. BIOVIA, BIOVIA Discovery Studio 2017 R2: a Comprehensive Predictive Science Application for the Life Sciences, in, San Diego, (2017).
[15] Schrödinger, Desmond molecular dynamics system, New York, Maestro-desmond Interoperability Tools, Schrödinger, 2016.
[16] H. Wickham, ggplot2, New York, Springer-Verlag, New York, 2009, p. 213.
[17] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[18] I.F. Fokkema, P.E. Taschner, G.C. Schaafsma, J. Celli, J.F. Laros, J.T. den Dunnen,

LOVD v.2.0: the next generation in gene variant databases, Hum. Mutat. 32 (2011) 557–563.

[19] S. Leigh, M. Futema, R. Whittall, A. Taylor-Beadling, M. Williams, J.T. den Dunnen, et al., The UCL low-density lipoprotein receptor gene variant database: pathogenicity update, J. Med. Genet. 54 (2017) 217–223.

[20] H. Wang, S. Xu, L. Sun, X. Pan, S. Yang, L. Wang, Functional characterization of two low-density lipoprotein receptor gene mutations in two Chinese patients with familial hypercholesterolemia, PLoS One 9 (2014) e92703.

[21] L. Wang, J. Lin, S. Liu, S. Cao, J. Liu, Q. Yong, et al., Mutations in the LDL receptor gene in four Chinese homozygous familial hypercholesterolemia phenotype patients, Nutr. Metabol. Cardiovasc. Dis. 19 (2009) 391–400.

[22] M. Walus-Miarka, M. Sanak, B. Idzior-Walus, P. Miarka, P. Witek, M.T. Malecki, et al., A novel mutation (Cys308Phe) of the LDL receptor gene in families from the South-Eastern part of Poland, Mol. Biol. Rep. 39 (2012) 5181–5186.

[23] A. Etxebarria, A. Benito-Vicente, A.C. Alves, H. Ostolaza, M. Bourbon, C. Martin, Advantages and versatility of fluorescence-based methodology to characterize the functionality of LDLR and class mutation assignment, PLoS One 9 (2014) e112677.

[24] D.W. Russell, M.S. Brown, J.L. Goldstein, Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins, J. Biol. Chem. 264 (1989) 21682–21688.

[25] E.J. Boswell, H. Jeon, S.C. Blacklow, A.K. Downing, Global defects in the expression and function of the low density lipoprotein receptor (LDLR) associated with two familial hypercholesterolemia mutations resulting in misfolding of the LDLR epidermal growth factor-AB pair, J. Biol. Chem. 279 (2004) 30611–30621.

[26] Y. Gao, H. Yin, Y. He, J. Wu, S. Wang, W. Li, et al., Prevalence and Outcomes of Familial Hypercholesterolemia Patients in a Chinese Myocardial Infarction Cohort vol. 2, Atheroscler Open Access, 2017.

[27] M.D. Di Taranto, M.N. D'Agostino, G. Fortunato, Functional characterization of mutant genes associated with autosomal dominant familial hypercholesterolemia: integration and evolution of genetic diagnosis, Nutr. Metabol. Cardiovasc. Dis. 25 (2015) 979–987.

[28] S. Calandra, P. Tarugi, H.E. Speedy, A.F. Dean, S. Bertolini, C.C. Shoulders, Mechanisms and genetic determinants regulating sterol absorption, circulating LDL levels, and sterol elimination: implications for classification and disease risk, J. Lipid Res. 52 (2011) 1885–1926.

[29] Y. Choi, A.P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (2015) 2745–2747.

[30] J.H. Chang, J.P. Pan, D.Y. Tai, A.C. Huang, P.H. Li, H.L. Ho, et al., Identification and characterization of LDL receptor gene mutations in hyperlipidemic Chinese, J. Lipid Res. 44 (2003) 1850–1858.

[31] J.B. Dube, J. Wang, H. Cao, A.D. McIntyre, C.T. Johansen, S.E. Hopkins, et al., Common low-density lipoprotein receptor p.G116S variant has a large effect on plasma low-density lipoprotein cholesterol in circumpolar inuit populations, Circ. Cardiovasc. Genet. 8 (2015) 100–105.

[32] F. Gao, H.E. Ihn, M.W. Medina, R.M. Krauss, A common polymorphism in the LDL receptor gene has multiple effects on LDL receptor function, Hum. Mol. Genet. 22 (2013) 1424–1431.