OXFORD

Human Molecular Genetics, 2016, Vol. 25, No. 6 1233–1246

doi: 10.1093/hmg/ddw004
Advance Access Publication Date: 10 January 2016
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: linking SNPs with disease phenotypes in familial hypercholesterolemia

Vladimir Espinosa Angarica[1,2,†], Modesto Orozco[3,4,5] and Javier Sancho[1,2,6,*]

[1]Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain, [2]Biocomputation and Complex Systems Physics Institute (BIFI), Joint Unit BIFI-IQFR (CSIC), Universidad de Zaragoza, Mariano Esquillor, Edificio I + D, 50018 Zaragoza, Spain, [3]Institut de Recerca Biomèdica (IRB Barcelona), Baldiri Reixac 10-12, 08028 Barcelona, Spain, [4]Departament de Bioquímica i Biologia Molecular, Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Spain, [5]Joint BSC-CRG-IRB Program in Computational Biology, Baldiri Reixac 10-12, 08028 Barcelona, Spain, and [6]Aragon Institute for Health Research (IIS Aragón), Universidad de Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain

*To whom correspondence should be addressed at: Biocomputation and Complex Systems Physics Institute (BIFI), Mariano Esquillor, Edificio I + D, 50018 Zaragoza, Spain. Tel: +34 976761286; Fax: +34 976762123; Email: jsancho@unizar.es

Abstract

Familial hypercholesterolemia (FH), a genetic disorder with a prevalence of 0.2%, represents a high-risk factor to develop cardiovascular and cerebrovascular diseases. The majority and most severe FH cases are associated to mutations in the receptor for low-density lipoproteins receptor (LDL-r), but the molecular basis explaining the connection between mutation and phenotype is often unknown, which hinders early diagnosis and treatment of the disease. We have used atomistic simulations to explore the complete SNP mutational space (227 mutants) of the LA5 repeat, the key domain for interacting with LDL that is coded in the exon concentrating the highest number of mutations. Four clusters of mutants of different stability have been identified. The majority of the 50 FH known mutations (33) appear distributed in the unstable clusters, i.e. loss of conformational stability explains two-third of FH phenotypes. However, one-third of FH phenotypes (17 mutations) do not destabilize the LR5 repeat. Combining our simulations with available structural data from different laboratories, we have defined a consensus-binding site for the interaction of the LA5 repeat with LDL-r partner proteins and have found that most (16) of the 17 stable FH mutations occur at binding site residues. Thus, LA5-associated FH arises from mutations that cause either the loss of stability or a decrease in domain's-binding affinity. Based on this finding, we propose the likely phenotype of each possible SNP in the LA5 repeat and outline a procedure to make a full computational diagnosis for FH.

†Present address: Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. 6, Avenue du Swing, L-4366 Esch-sur-Alzette, Luxembourg.
Received: September 29, 2015. **Revised:** December 19, 2015. **Accepted:** January 5, 2016

Introduction

The low-density lipoprotein (LDL) receptor (LDL-r) belongs and gives name to an ancient family of membrane receptors, including very-low-density lipoprotein receptor (VLDL-r), ApoER2, LRP1, LRP2 and LRP6 (1), that appeared early with the onset of the first metazoans and play important roles in multiple biological processes, through binding to a diverse set of partners (2,3). The receptors of the LDL-r family contain a common set of structural constituents that, from C-terminal to N-terminal, include (a) a cytoplasmic region that embodies *NPxY* and *PPPSP* motifs, (b) a single-transmembrane segment anchoring the cytoplasmic and extracellular sections to the cell membrane and (c) an extracellular region formed by an epidermal growth factor (EGF)-like domain composed by several EGF repeats and a β-propeller domain, followed by a ligand-binding region consisting of a variable number of small cysteine-rich domains (1,3), known as LDL-r type A domains (LA domains). The ligand-binding region of the human LDL-r (4) has been widely studied to uncover the mechanism of endocytic LDL internalization and release. These studies have provided a wealth of structural data corresponding to individual domains, domain pairs, the complete extracellular region (5–8) and even a low-resolution structure of the LDL–LDL-r complex (9). Domains LA1–7 (10) and, most importantly, domains LA4–5 (2,11), are key for binding of VLDL and LDL particles (12). LA domains are 40-residue, small autonomous folding units containing little secondary structure and lacking an extensive hydrophobic core, which are mainly stabilized by three disulfide bridges and a coordinated calcium ion (2,13–16). The seven LA domains are connected through small peptide linkers that provide great flexibility to the region (8).

Familial hypercholesterolemia (FH) is a genetic disorder associated to abnormally high levels of LDLs in the blood, which constitute a significant-risk factor for cardiovascular and cerebrovascular diseases (17–20), and has a prevalence of 1:500 in heterozygosis in human populations (21,22). Although FH can be caused by defects in several proteins linked to cholesterol internalization and metabolism in cells—e.g. Apo B-100 (23,24), PCSK9 (25,26) and the LDL-r (27,28)—the majority and most severe FH cases are associated to mutations in the LDL-r (29), some of which have been shown to compromise the conformational stability of specific receptor domains (30). In spite of the high-prevalence worldwide, FH is under-diagnosed and undertreated (31), probably because of the complexity for connecting genetic variations and disease phenotype. Since the recognition of the association of FH to genetic variations in the LDL-r and the discovery of the first disease-causing mutations (32–34), much information has been gathered on different types of mutations in the LDL-r gene (28,35–38). The number of known mutations for the LDL-r is between 1741 and 1835, according to current releases of the LDL-r database (28) and the Professional version of the human gene mutation database (37), respectively. Genetic variations found in this protein include large rearrangements of coding and/or intronic regions, synonymous and non-synonymous substitutions and mutations in the regulatory regions or splicing sites (28,37). Missense substitutions are by large the most frequent type of mutation (28), and are unevenly distributed along the LDL-r gene sequence. A higher accumulation of genetic variations has been reported in exons coding for the ligand-binding region, particularly in exon 4, coding for LA domains 3–5 that are key for LDL binding (2,11,12).

The development of high-scale DNA genotyping and sequencing methodologies (39–42) has stimulated an increase of cascade screening programs in partial populations of some countries and high-risk groups (40–44). Unfortunately, due to a lack of knowledge on the phenotypic effect of most mutations, standard genotype analyses are centered only in a reduced set of known pathological mutations, limiting the predictive power of these techniques (45). Furthermore, the lack of understanding of the molecular basis of the pathological effect of LDL-r mutations limits our ability for devising novel therapies for treating FH. To gain such molecular insight *in vitro* and *in silico* studies have been performed on different domains or on the complete LDL-r-binding region (2,5–7,13–16,46–49). These studies provide insights to relate the severity of mutations with structural or stability impairments in the LDL-r. However, we are still far from having a complete molecular-level description of the connection between genetic variations in the LDL-r gene and the severity of FH phenotypes.

In this work, we have performed massive atomistic simulations for predicting the fate of all possible missense mutations in the key LA5 LDL-r lipoprotein-binding domain. Thus, we have generated all the possible mutants arising from non-synonymous single nucleotide polymorphism (SNPs)—i.e. 256 SNPs coding for 227 different mutants (Supplementary Material, Fig. S1)—and have run MD simulations to assess the distortions caused by single-amino acid substitutions. When the structural fluctuations observed during MD trajectories are moderate, we classify the corresponding mutations as not destabilizing. Conversely, they are referred to as destabilizing when they cause a significant distortion of the LA5 domain structure. However, the data we use for classifying mutants in one of these two categories should not be considered as quantitative estimations of changes in the thermodynamic stability of the protein, which is not measured in this study. A total of 4.5 microseconds of MD simulations corresponding to relaxation trajectories have been analyzed using a variety of data-mining methodologies. The use of these analyses, together with our proposal of a consensus-binding site of the LA5 domain based on structural information obtained in different laboratories (8,50), allows us to explain the pathological nature of most mutations described for the LA5 domain as either arising from structural destabilization of its tridimensional structure (Supplementary Material, Fig. S2), or from replacement of binding site residues. The results presented here provide a simple approach for *ab initio* prediction of the putative pathological impact of new mutations, opening a way for an early diagnosis and treatment of FH.

Results

Global conformational instability of the LA5 domain mutant variants

The 227 mutants resulting from SNPs in the LA5 domain (Supplementary Material, Fig. S1 and Table S2) were generated *in silico* using SCWRL (51), which was also used to find the best rotamer conformations for the side chains of the substituted residues. All mutants were minimized and equilibrated in explicit water and MD simulations were extended for a total aggregated time of 6 μs, including preparation and production trajectories (see the 'Materials and Methods' section). For each mutant 20 ns-long MD simulations were used to explore its conformational evolution upon mutation. Based on previous work (49), this time span was considered enough to permit significant relaxation from the initial wild-type-like structure. A parameter commonly used to evaluate the structural similarity of two proteins, or of two protein conformations along an MD trajectory, is the root mean square deviation (RMSD) or average distance between the atoms of the superimposed structures. However, some

characteristics of the RMSD (it is length-sensitive and it is not normalized) are not ideal to measure structural similarity. On the other hand, the template modeling score (TM-score) (52,53), constitutes a protein-length independent metric, which allows performing thorough structural comparison for identifying topology relationships among related proteins. It has been demonstrated, from a comprehensive comparison of proteins from different folding categories, that TM-scores close to 1 correspond to proteins belonging to the same fold topology, while for proteins with different structural topologies values <0.5 are obtained (52). We, thus, followed the dynamical evolution of different mutants using the TM-score. The evolution of the TM-scores for a selection of the 227 trajectories corresponding to some of the mutations found in individuals with FH is shown in Supplementary Material, Figure S3. The extent of the conformational change varies greatly among mutations. For example, the substitutions C197(176)Y; F200(179)L, C; C204(183)S, Y; S206(185)R; D221(200)Y and D227(206)V (see Supplementary Material, Tables S1 or S2 for details on the numbering) cause great conformational change in this timescale leading to conformations containing significant structural distortions after a few nanoseconds of simulation. In contrast, other FH known mutations such as S198(177)L; C209 (188)Y; H211(190)D; W214(193)S; D224(203)G; C231(210)R and C231(210)Y cause only mild or apparently no distortions to the structure of the LA5 domain during the simulations. To evaluate the degree of conformational distortion associated to each mutant we have performed principal component analysis (PCA) of all trajectories (54–56). PCA is a standard statistical procedure (see the 'Materials and Methods' section) for performing transformations of multivariate data for identifying correlation among variables in the initial data set. This approach has been widely used to analyze MD simulation data (54,57), allowing the decomposition of the trajectories into simple uncorrelated motions and the identification of the more relevant ones. Those essential motions (principal components) constitute highly compressed representations of entire trajectories and offer convenient ways to visually represent and compare different trajectories. When dealing with large numbers of long trajectories, analysis and comparison using principal components is much easier and more objective than using conventional structural analyses of individual trajectories. A summary of the analysis of the 227 mutant trajectories is included in Supplementary Material, Figure S4. The 'Scree Plot' in this figure for the 1st to 30th components shows that the first three ones describe on average 36, 16 and 9% of the total variance (i.e. they amount for >60% of the explored MD variance). Those three eigenvectors have been selected to describe the 'essential dynamics' of the systems (54,58–61).

After performing the PCs decomposition of all trajectories, we have made a pair-wise comparison of the average structures of different mutants. As shown in Figure 1, while many mutants display fairly similar average structures, some groups of mutants are structurally dissimilar to most others, e.g. trajectories M009–M010, M029, M049–M053, M153–M155, M161–M164 and M191–M197 (see Supplementary Material, Table S2 for the correspondence among trajectory indexes, codon change and amino acid substitutions), The conformational evolution of some of these dissimilar mutants along the trajectories, as shown by the projections of each trajectory into the first three PCs, is depicted in Figure 2, Supplementary Material, Figure S5 and Videos S1–S5. The charts in Figure 2 show the projections of the conformations visited during a simulation into the space formed by the first three PCs. The values of the projections provide a measure of the similarity of the structure at a given time step with the average structure of the simulation, located at the origin $(0_{1^{st}}, 0_{2^{nd}}, 0_{3^{rd}})$. For a Boltzmann-weighted ensemble of molecules that behave harmonically, a 'Gaussian' distribution of projections in each dimension is expected, which correspond to an ellipsoid centered at the origin, with the eigenvectors corresponding to the axes of the plot, and the dispersion of the ellipsoid length in the ith PC being proportional to the ith eigenvalue. Deviations from this expected behavior could be taken as a qualitative estimate of the conformational change caused by the mutation. The significant conformational change caused by some amino acid substitutions on the LDL-r LA5 structure is clear when Figure 2 (and Supplementary Material, Fig. S5 and Videos S1–S5) is compared with the plots corresponding to background mutants (Fig. 3, Supplementary Material, Fig. S6 and Videos S6–S10) that has little or no effect on the conformation of the LDL-r LA5 domain.

Clustering of LA5 mutants according to the extent of conformational instability

Among the 227 trajectories collected some are stable, but there are also unstable trajectories that do not follow a multivariate normal distribution in the PC space. In order to perform a combined analysis of all the trajectories, we have concatenated the last 10 ns of each one into a meta-trajectory. This provides a common PC space and Eigensystem where the independent trajectories can be compared. Within this new reference system, we have used a sampling methodology to compare the essential subspaces of the trajectories. By randomly comparing subsets of each simulation against each other (10^5 random comparisons), it was possible to obtain a statistical assessment of the mean distance between the essential subspaces visited by the different mutants and the wild-type LA5 domain. The Mahalanobis distance metric (62) was used for comparing and clustering the mutants according to the extent of conformational instability (see the 'Materials and Methods' section and Fig. 4), which allowed to objectively establish a link between conformational instability and the likelihood of the expression of a LDL-r variant with an impaired function towards LDLs interaction (2,11,49).

The results in Figure 4 provide an intuitive picture of the clustering of the trajectories in a 25th-dimensional space, sufficient to describe 95% of the conformational variability. As expected, the projections for the trajectories of the wild-type LDL-r LA5 domain and of different stable mutants explored an ellipsoidal subspace close to the PCs origin, and consequently they formed a stable cluster of mutants shown in green. This is the largest cluster, comprising 114 mutants (for a detailed list of the mutants see the color codes in Supplementary Material, Table S2). Three additional clusters appeared that are formed by 57 unstable mutants (orange cluster), 34 very unstable mutants (magenta cluster) and 22 highly unstable mutants (red cluster), as judged from the conformational changes experienced. So far, 50 disease-linked SNP have been found in the LDL-r LA5 domain of FH individuals (28,37). These pathological mutations appear unequally distributed among the four stability clusters. The large stable cluster only includes 17 (34%) of the 50 known FH mutations, the remaining 33 being distributed among the three unstable clusters (Fig. 4). The stability of the LA5 structure in the 17 known diseased-linked SNPs classified by us as 'stable' mutations can be appreciated in Supplementary Material, Figure S7 and Videos S11–S15. We observe two trends in the structural distribution of mutants among the four clusters. The percentage of buried mutations increases from 52% in the stable cluster to 72, 88 and 86% in the progressively more unstable clusters, and a similar increase is observed in

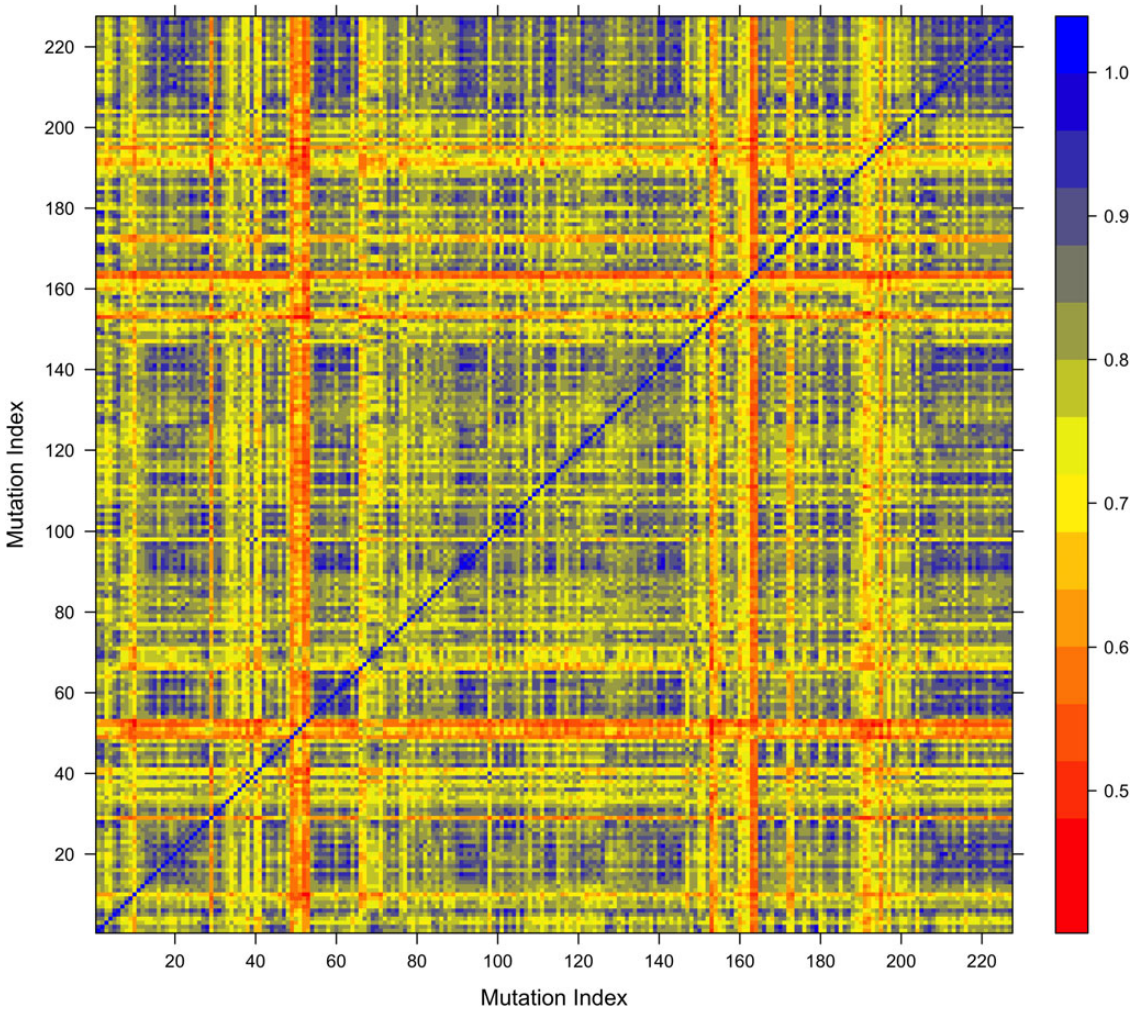

Figure 1. All-to-all comparison of the trajectories average structures. The average structures of all the 227 mutants were extracted from their corresponding trajectories after performing a PCA. The average structures were compared in pairs using the TM-score metric. The chart is a heat map of the comparison of all versus all, and the color in each cell corresponds to the TM-score for the comparison of two structures whose indexes are found in the abscissa and ordinate axes. The rightmost side of the chart shows the color legend for the TM-score, from red for dissimilar structures (TM-score ≈ 0.5) to blue for identical structures (TM-score ≈ 1).

mutations affecting Ca^{++} coordinating residues or cysteines (from 25 to 25, 50 and 63%). It appears that, in agreement with structural/energetic expectations, mutations in buried residues and in those contributing to structural *loci* are, on average, more destabilizing that others.

Discussion

Direct assessment of the effect of all possible SNPs in the structural stability of the LDL-r LA5 domain

Conformational diseases (63–66) are pathological states arising from alterations in protein conformational or binding equilibria driven by changes in physicochemical conditions or by mutations. In FH, an ample number of mutations identified by cascade screening assays (40,41,43,44) have been reported to be genetic determinants of the disease. Although there have been attempts to experimentally assess the effect of mutations in some LDL-r domains (5–7,13–15,46–48), we are far from being able to experimentally explore the consequences of all the biologically accessible mutations. Only a small fraction of all possible mutations in the LDL-r has been catalogued and reported in genetic variation and sequence databases (28,35–38). In an effort to fill the gap between possible genetic variability and knowledge of the associated phenotype, many computational methodologies have been developed (67,68). Those methods, mainly based on indirect genetic, structural or evolutionary assumptions, cannot always anticipate the real effect of the amino acid substitution at the structural level, a key information for predicting whether the mutation might cause a perturbation in the protein conformational equilibrium, and also a crucial step to derive structural information for more efficient drug-design protocols. We present here an alternative '*de novo*' computational strategy based on the analysis of relatively short, all-atom MD simulations of the protein under physiological conditions. This is to our knowledge, the first complete exploration of the effect of all biologically accessible mutations caused by SNPs in the structure of a protein domain: the LDL-r LA5-binding domain, to determine whether and how FH might be etiologically related to genetic variations.

We have generated all possible mutants arising from SNPs in the cDNA for the LDL-r LA5-binding domain (Supplementary Material, Fig. S1), performing for all of them atomistic MD simulations under physiological conditions. In this 'exhaustive' approach, instead of concentrating on explaining previously

identified specific amino acid substitutions, we have explored the entire SNP mutational landscape of a key domain known to establish functional interactions with LDLs (2,11), and encoded in the LDL-r exon bearing the highest proportion of mutations identified in individuals with FH. Inspection of the 3D structure of the LA5 domain (Supplementary Material, Fig. S2) allows identifying important structural *loci* on which the substitution of an amino acid could be accompanied by significant conformational changes. However, our results point out to a context-dependent scenario where the specific substitution, not just the *locus*, determines whether the structure of the LA5 domain is significantly affected. This finding is illustrated in Supplementary Material, Figure S3 with a selection of trajectories corresponding to FH mutants (28,37). Their TM-scores (52,53) prove that, even in mutants involving

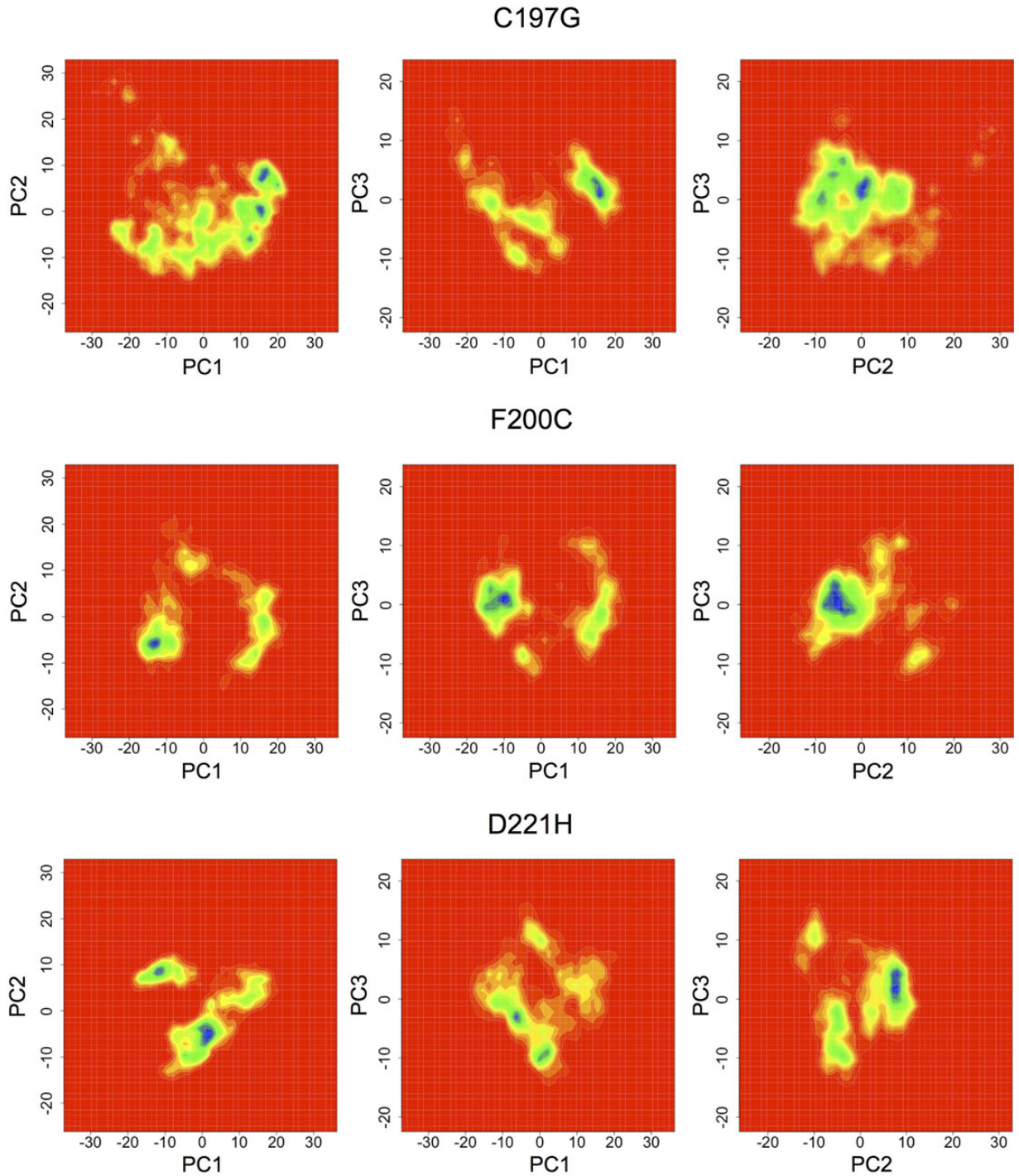

Figure 2. Dynamical evolution of LA5 mutants in the PCA space (destabilizing mutations). The MD trajectories are followed along time by projecting the structures at each time step into the space described by the first three PCs. Each subchart is a two-dimensional density plot of the projections of the structures into PC1 versus PC2, PC1 versus PC3 and PC2 versus PC3. The color scale goes from red (no occupancy) to blue (high occupancy), passing through intermediate scales of yellow and green. For accessing the more descriptive animations please visit the corresponding files for each simulation in the Supplementary Material, Videos S1–S5.

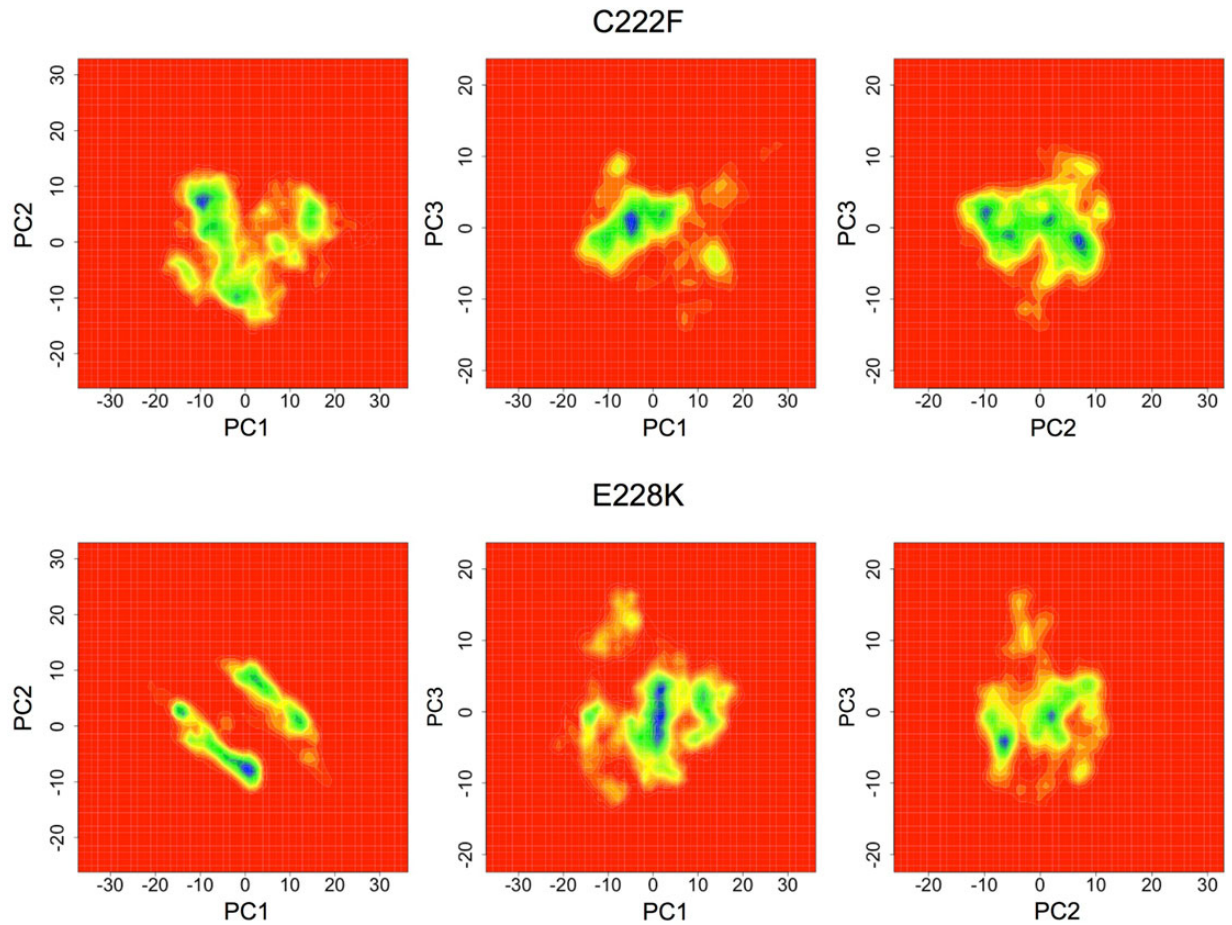

Figure 2. Continued

residues from the calcium coordinating box or disulfide bond-forming cysteines, some simulations show stable evolutions along time while, for others, the conformational stability appears to be significantly affected. These results are not obvious and could not be easily predicted by standard trained methods (67–71).

In order to define quantitative criteria for a global comparison of the conformations visited by different types of mutants during MD simulations we performed PCA of each trajectory, and the structural differences among the corresponding average structures were calculated using the TM-score (Fig. 1). The presence of clusters of mutants structurally different from the rest of variants suggests the possibility of grouping mutants according to the extent of the instability introduced by the amino acid substitution. The PC representations depicted in Figure 2, corresponding to destabilizing mutants most of whom are associated to FH (Supplementary Material, Table S2), graphically confirms the instability caused by these mutations. In contrast, Figure 3 shows several examples of stable evolutions around the average structure, including mutations such as C209(188)W in an important structural *locus*, which reaffirms the need for specific mutation testing versus general predictions based on structural location.

FH mutations in the LDL-r LA5 domain due to loss of conformational stability or binding competence

A main goal of this work is to devise a quantitative way of identifying SNP mutations that could destabilize the structure of the LDL-r LA5-binding domain and to group different types of mutations according to the extent of destabilizing effects. Analysis of the meta-trajectories of the 227 mutants allows comparing, within a single Eigensystem, the conformational subspaces more probably visited by each mutant. The clusters in Figure 4 provide an objective classification of the SNPs according to their effect in structural stability, and hence their possible pathogenicity (see also in Supplementary Material, Table S2 a color-coded classification of the destabilizing effect of each mutant). Of the 50 known FH mutations, 33 appear distributed in the three unstable clusters, indicating that loss of conformational stability explains two-third of FH phenotypes. According to our simulations, there appear to be hotspots in the structure of the LDL-r LA5 domain where SNPs are more likely to lead to substitutions compromising the conformational stability—e.g. C197, E201, C204, D221, C222, D224, D227 and E228 (Supplementary Material, Fig. S8). Conversely, some positions are unlikely to derive in unstable mutant proteins due to SNP—e.g. A199, L205, G219, N230 or A232.

On the other hand, 17 FH mutations (Supplementary Material, Table S2) are classified as stable, and the individual analysis of their trajectories (see examples in Supplementary Material, Fig. S7) confirms they do not significantly alter the conformational behavior of the LA5 domain. This is a clear indication that they will cause FH by a different mechanism. The simplest reason that can explain their relationship with FH is that those mutations impair the interaction of LA5 with its binding partners, either other domains from the LDL-r or other proteins involved in cholesterol homeostasis. There are three known partners of LA5: the β-propeller domain of the LDL-r, and the apolipoproteins Apo

B and Apo E present in LDL and/or VLDL particles. At low pH (8), LA5 interacts with the β-propeller domain through residues H211, S212, W214, D217, G219, D221 and K223, clustered in the LA5 convex face (Fig. 5). Recent structural data from our group (50) have determined that the interaction between LA5 and key interacting helices of Apo E and Apo B also involves essentially residues at the convex face (W214, G218, G219, D221 and D227 in the complexes with Apo B and Apo E, plus E201 in the complex with Apo E). Those two sets of residues define a common patch in the convex face (Fig. 6A) that very likely constitutes the binding site used by LA5 to interact with different partners. In fact, the homologous LDL-r domains LA3 and LA4 use a structurally equivalent patch to recognize yet another partner, the receptor-associated protein (RAP) (6).

The 17 FH mutations in the stable cluster occur in 11 residues (P196, S198, E201, C209, H211, W214, C216, D221, D227, E228 and

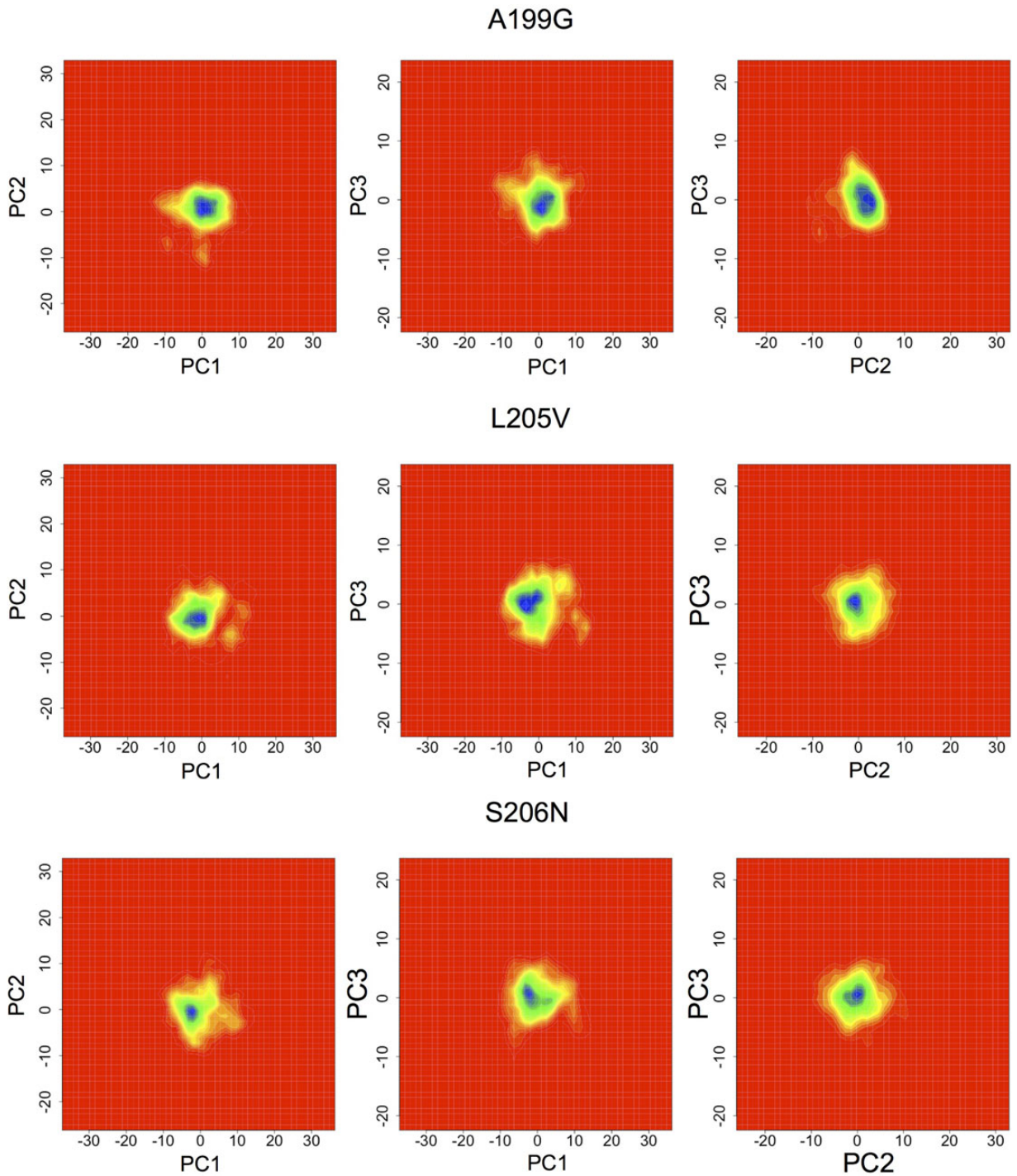

Figure 3. Dynamical evolution of LA5 mutants in the PCA space (non-destabilizing mutations). The MD trajectories are followed along time by projecting the structures at each time step into the space described by the first three PCs. Each subchart is a two-dimensional density plot of the projections of the structures into PC1 versus PC2, PC1 versus PC3 and PC2 versus PC3. The color scale goes from red (no occupancy) to blue (high occupancy), passing through intermediate scales of yellow and green. For accessing the more descriptive animations please visit the corresponding files for each simulation in the Supplementary Material, Videos S6–S10.

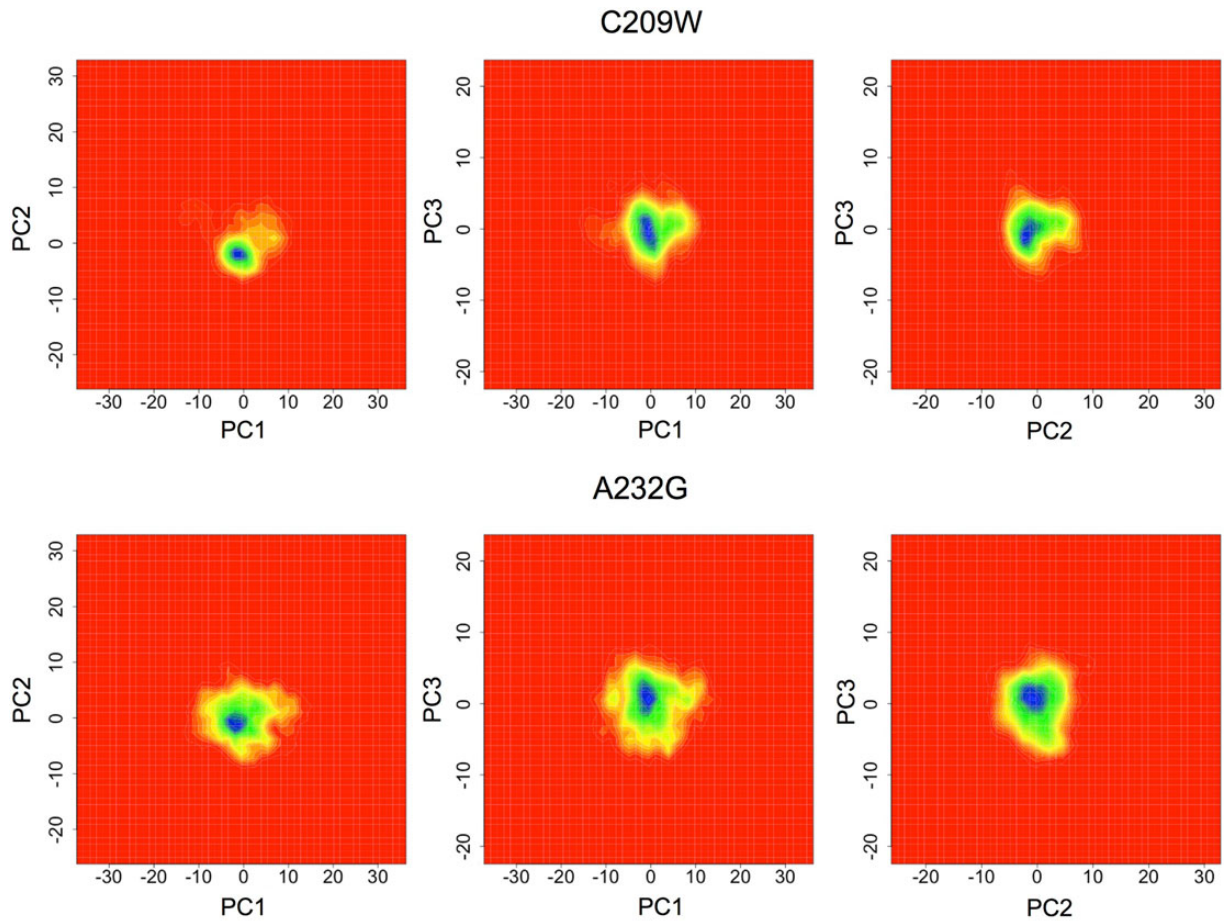

Figure 3. Continued

C231), 7 of which are surface exposed and contiguous, defining an extended but overlapping version of the binding site (Fig. 6B). Three of the four remaining residues, C216 and C231 forming the C-terminal most disulfide bridge, plus E228 at the calcium cage, are also surface exposed at one end of the convex face and appear to constitute a prolongation of the known binding site (Fig. 6B). Only D227 is buried. As expected, most substitutions at this residue are greatly destabilizing (Supplementary Material, Table S2), but its replacement by similarly charged glutamic acid, although reported to cause FH, does not lead to structural perturbations during the simulations. The likely cause of the FH character of D227(206)E is that its reduced Ca^{++} affinity impairs folding (13). Except for this mutation, the rest of the 17 non-destabilizing FH SNPs in the LA5 domain affect binding site residues. Thus, the pathogenicity of these mutations seems related to disruption of LA5-binding compatibility with other proteins. The underlying structural reasons can be either that the substitution abolishes important contacts with partner-binding residues—e.g. W214 (193)S (8)—or that the newly introduced residue modifies the steric or physicochemical compatibility of the interacting patches [e.g. H211(190)D, Y, L (8,72)].

From molecular dynamics to a strategy for computational diagnosis in conformational diseases

Our results indicate that 33 out of the 50 known FH mutations at the LA5 domain are destabilizing and that, of the remaining 17 non-destabilizing mutations, 16 occur in residues at the surface-binding site. As it appears, computational estimations of conformational instability from short-range MD simulations, combined with experimental knowledge of the LA5 domain interacting residues, makes possible to anticipate the disease phenotype that has been observed in 49 of the 50 SNPs known to be related to FH. This result shows that our method has a remarkably high sensitivity (true positives rate) of 0.98 for classifying FH-causing SNPs. We have also tried to provide a measure for the specificity of the method (true negatives rate), but the lack of sequence data on neutral mutations (see legend of Supplementary Material, Table S2) impedes it.

We would like to propose a FH computational diagnosis for all possible SNPs in module LA5 that considers both the degree of conformational instability in the simulations and the possible impairment of the interacting region (Supplementary Material, Table S2). Of all the possible SNPs—i.e. 256 non-synonymous SNPs coding for 227 different single amino acid substituted variants—only 22% have been found in FH individuals (dot tagged mutants in Supplementary Material, Fig. S1). These variants are labeled as deleterious in Supplementary Material, Table S2. For the remaining 177 mutations not reported in genetic variation databases, those belonging to any of the three unstable clusters are also classified as deleterious. For those in the stable cluster, we have considered evaluating their possible binding impairing effects using qualitative structural criteria—e.g. steric and physicochemical differences between the wild-type and substituted residues. However, such a fine evaluation would be too preliminary given the limited structural information available for

LA5/partner complexes, e.g. the structure of the LDL-r complete extracellular region (8) is of low resolution, and the interactions of LA5 with Apo B and E peptides are only defined in part (50)

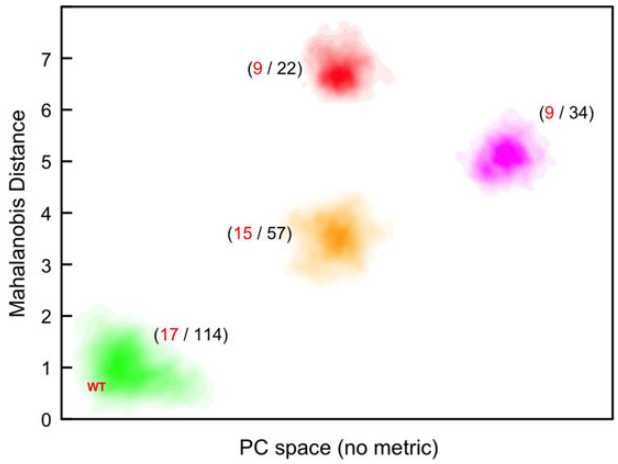

Figure 4. Clustering of LA5 mutants according to the extent of conformational instability. For the meta-trajectory of the concatenation of the last 10 ns of the 227 simulations, the average Mahalanobis distance among all pairs of simulations was used to assess the difference in the subspaces explored by each mutant in the PCA N-dimensional space (25-dimensions). Based on those distances, a complete-link-based clustering algorithm was used, and an abstract representation of the four more representative clusters is shown. The clusters are color-coded: green (stable mutants), orange (unstable mutants), magenta (very unstable mutants) and red (highly unstable mutants). For each cluster, we show in parenthesis the number of mutants found in persons with FH (in red) and the total number of mutants. The dispersion observed in each cluster corresponds to the variability observed for the average Mahalanobis distance of each simulation and the rest of the simulations included in the corresponding cluster, which correspond to branching nodes representing these trajectories in the clustering dendrogram.

because the exact conformation of those peptides in the complexes is not known. Therefore, we have provisionally evaluated all mutations taking place in binding site residues as 'deleterious', which may increase the number of false-positives in this subset of our predictions. Our phenotype predictions in Supplementary Material, Table S2 can be compared with predictions calculated using different methodologies, such as PMUT (68), and a consensus approach, CONDEL (69), integrating the predictions made using SIFT (67,73), polyphen-2 (71) and mutation assessor (70,74). We have also included the predictions obtained using polyphen-2 (71), and the calculation of stability changes upon mutation obtained with FoldX (75). The comparison reveals clear discrepancies among the different predictions, and stability estimations in key structural *loci*, such as in cysteines or in Ca^{++}-binding residues—e.g. in C197(176)S or E228(207)D as well as in many other mutations in the LA5 structure (Supplementary Material, Tables S2 and S3). Thus, depending on the predictive approach used, the conclusions drawn would be different. Moreover, the true positive rates obtained with PMUT, CONDEL and polyphen-2 for the classification of FH-causing mutations are 42, 76 and 80%, respectively (Supplementary Material, Tables S2 and S3), which shows that our structure-based method outperforms all these sequence-based approaches. Furthermore, we are not only able to correctly predict almost all FH-causing mutations, but also to differentiate mutations that cause the disease through the structural instability of the LA5 domain, and others associated to residues in the interaction site with other partner proteins and LDL particles. Though undoubtedly our approach is more computationally expensive and requires more data processing and analysis than others available for predicting deleteriousness of mutations (67–71,73,74), general advances in computation speed and specific improvement in MD simulations (76–78), together with the emergence of online services for performing client-based and high-throughput MD simulations (79–81), may facilitate the generalization of the approach presented here in the near future.

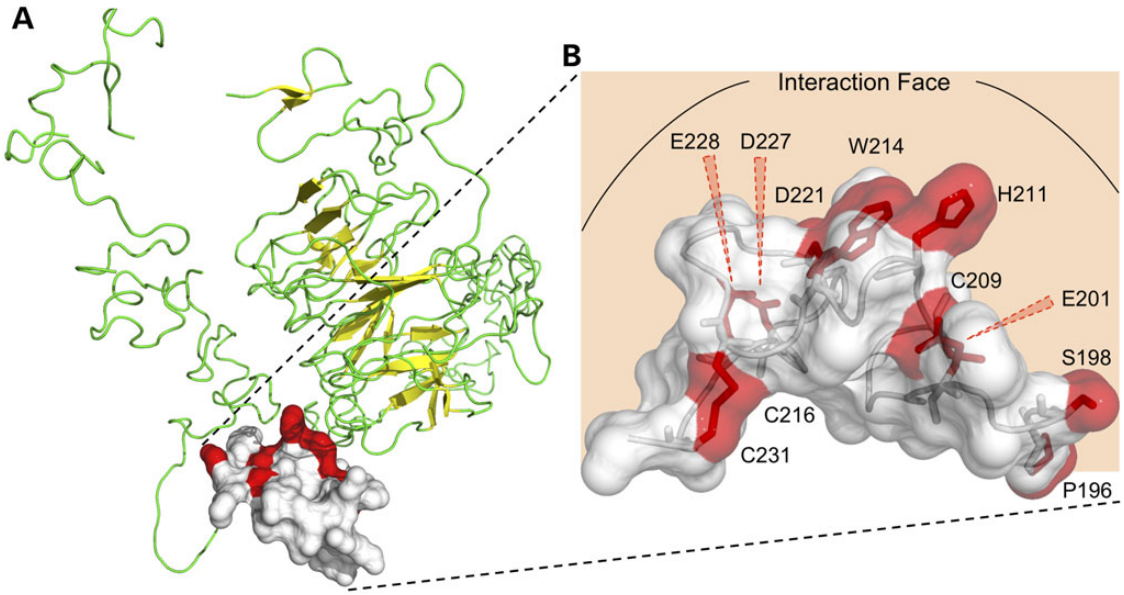

Figure 5. The binding region of the LDL-r LA5 domain. The structure of the LDL-r LA5 domain and the interaction region. (**A**) The LA5 domain in the context of the structure of the complete LDL-r extracellular region (PDB id: 1N7D). The LA5 domain is shown in surface representation colored in white, highlighting in red the 11 residues where the 17 mutations not affecting the conformational stability of the domain occur. (**B**) A close look of the LA5 domain and the 11 residues bearing FH mutations that do not destabilize the domain. We highlight the upper convex region of the LA5 domain, which according to recent experimental evidence (50), is responsible for the interaction of the domain with LDL and VLDL particles, in addition to forming the self complex shown in (A). From the residues highlighted in red, we also include their name and position in the sequence. Of the 11 residues showed, only residue D227 is buried.

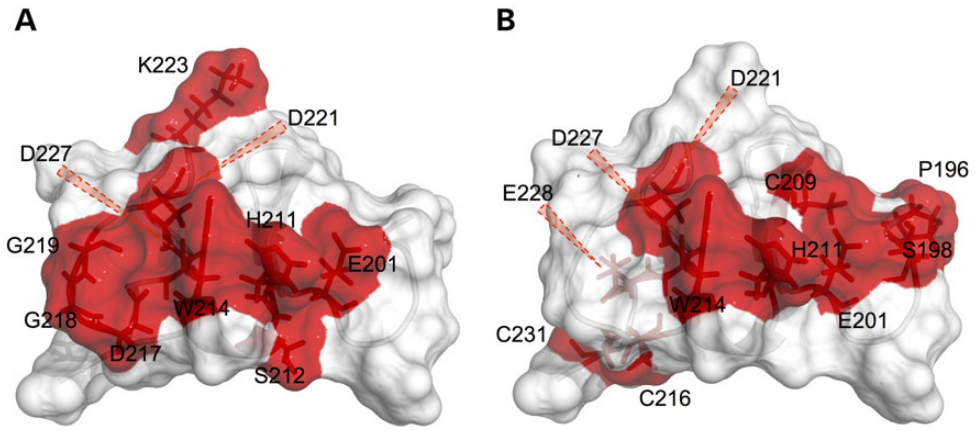

Figure 6. The consensus-binding region of the LDL-r LA5 domain compared with the region bearing non-destabilizing FH mutations. (**A**) The residues in the convex face of the LDL-r LA5 domain known to participate in interactions with other domains from the LDL-r (8) or from Apo B and Apo E (50), are highlighted. (**B**) The residues that bear some mutations related to FH that do not destabilize the domain and are classified in the stable cluster, are highlighted. Ten out of the 11 residues in which the 17 non-destabilizing known mutations are distributed at the surface, can be viewed simultaneously. Only residue D227 is buried.

Our analysis also indicates that, together with LDL-r destabilizing mutations, a significant percentage of FH phenotypes are related to impairment of structural regions mediating protein/protein interactions (e.g. LR5/β-propeller domain or LR5/apoB). In this respect, a complete picture of the interactions formed by the different domains of the LDL-r among them and with VLDL, LDL, PCSK9, RAP or additional partners yet to be discovered is lacking. Comprehensive high-throughput and high-resolution interactome structural predictions will eventually become available, but meanwhile much experimental work remains to be done. Besides, a structural and thermodynamic understanding of how the LDL-r interactions are affected by the specific and changing solution conditions in the different cellular compartments visited along its functional cycle is also needed (5–7,13–15,46–48,50,82–84). Although the structural distortions caused by mutations in the isolated LA5 domain, as identified with our approach, need not be identical to those taking place in the context of the whole LDL-r, the strong experimental evidence showing the structural independence of LA domains (8,30) suggests those distortions are likely to be very similar. Nevertheless, in cases where differences occur they will impose an upper limit to the predictive accuracy of the method. It is also important to emphasize that the evaluation and improvement of the specificity of this or any other predictive method will benefit from the search for, and documentation of, non-pathogenic mutations.

In this work, we have provided a clear structure/stability view of the complete mutational landscape of the LA5 domain, with a quantitative classification of the conformational instability caused by all biologically accessible SNP amino acid substitutions. We hope that these data would be useful for planning experimental work to measure the real extent of the structural instability associated to yet-to-study putative destabilizing mutations, and for designing screening devices for the efficient diagnosis of FH. By extension, the method here illustrated may be applied to studying how SNP may affect the structure and function of other proteins associated to other pathologies.

Conclusions

Examination on the entire SNP mutational space of the LDL-r LA5 domain using relaxation molecular dynamics simulations allows to accurately classify each possible mutation as either compatible with the native structure or as destabilizing. Comparison of this classification with the known SNPs associated to FH disease phenotype clearly reveals two types of FH mutations: those causing a stability defect and those impairing binding interactions with LDL-r-associated proteins. The data generated here delineate the space of putative pathogenic mutations in an important LDL-r domain and may help experimentalist to develop more comprehensive FH screening methods, and may contribute to a better understanding of FH from a structural perspective. On a larger scale and with sufficient computation power, it seems possible to make a full computational diagnosis for FH by considering both the degree of conformational instability in simulations of the LDL-r and related proteins, and the possible impairments of their interacting regions. Importantly, the structural approach followed by us can be applied to predict the deleteriousness of genetic variations in other small proteins without relying in the evolutionary assumptions characteristic of most current methods based on sequence analysis. An obvious challenge in applying this method to larger proteins is that they will require more computation power due to their larger size and also to the larger number of possible SNPs. An additional challenge may apply in the form of slower relaxation kinetics for large proteins (85), especially if they exhibit full thermodynamic cooperativity (86).

Materials and Methods

LA5 domain coding sequence, structure and complete SNP mutational map generation

From the protein sequence of the complete human LDL-r accessible in Uniprot (35) (ID: LDLR_HUMAN, AC: P01130), we extracted the DNA-coding sequence for the LA5 domain by accessing the entry for this gene in the Ensembl database (87) (ID: ENSG00000130164). The protein sequence for the LA5 domain corresponds to residues 195–233 in the sequence of the complete receptor, while the X-ray structure of the domain used as the starting point for MD simulations and structural analysis (PDB id: 1AJJ), includes residues 196–232. Thus, we just extracted the coding sequence for amino acids 196–232, leaving out the codons for the N- and C-terminal serine and valine. The DNA sequence was then processed with an *ad hoc* script for generating all the biologically accessible mutants arising from the substitution of a single

nucleotide (Supplementary Material, Fig. S1). All the non-synonymous SNPs were identified—i.e. 256 non-synonymous SNPs coding for 227 different single-residue substituted protein variants—and the corresponding mutations in the structure of LA5 domain were generated using the program SCWRL (51), for finding the best rotamers of the mutated residue. Then, the mutants were given a specific code (Supplementary Material, Table S2) to be further processed before running the MD simulations.

Setting up the systems for molecular dynamics simulation production

Each of the 227 mutants, plus the wild-type LA5 domain, was solvated in a cubic water box with approximately 5500 TIP3 water molecules, and neutralized with Na^+Cl^- counter ions using the solvate package in VMD (88). We setup a thorough procedure for preparing the systems previous to running the production MD simulations, including multiple cycles of step-descending minimization/equilibration steps in a preparation phase of ~5 ns, which encompasses: (a) short CPT dynamics of water molecules with the protein atoms fixed to eliminate possible potential strains in the water box, (b) slow release of the protein atoms by imposing decreasing elastic restraints and (c) very slow heating of the systems to the final simulation temperature (310K) using a gradient temperature ramp. The 5 ns preparation phase guaranteed the stabilization of the temperature and total energy of the systems. Then, 20 ns production MD simulations were run for each mutant using the CHARMM (89) force field (version c34b1) in NAMD (90). The simulations were run using Langevin dynamics, with periodic boundary conditions and Particle Mesh Ewald for modeling long-range electrostatic interactions with a cutoff distance of 14 Å. The Nosé–Hoover thermostat was used for pressure coupling of the system and the friction coefficients were set to 0.5 and 60 ps^{-1} for protein atoms, and water molecules and ions, respectively. The simulations were run mainly in the cluster of the Red Española de Supercomputación: *marenostrum* at the Barcelona Supercomputing Center and *CaesarAugusta* at the Institute for Biocomputation and Complex Systems Physics (BIFI), and also at the *Terminus* and *Memento* clusters at BIFI. The trajectories were analyzed with VMD (88) and a set of *ad hoc* TCL and Perl scripts.

Principal component analysis of individual MD trajectory data and of meta-trajectories

PCA, a useful procedure for capturing correlations among variables, has been extensively used for analyzing MD trajectories aiming at describing the 'essential dynamics' (54,58–61) of a system. Performing PCA on MD data starts by aligning the trajectory for removing the translational and rotational components of movement (91). Then, the trajectory is centered to the reference structure S_{ref}—e.g. the initial or an average structure—by subtracting the reference structure to the aligned snapshots, and it is represented as a matrix of the type $T_C = [3N \times F]$, on which the rows are the coordinates of the N residues of the system, and the columns the number of frames or snapshots, F, of the trajectory. Subsequently, the covariance or correlation matrix is calculated from the product of the trajectory matrix by its transpose $\sum = (1/3N)T_C \cdot T_C^T$. The eigenvalue decomposition of the covariance matrix renders a set of eigenvalues and orthogonal eigenvectors organized in the form $\Lambda = V^T \cdot \Sigma \cdot V$, where Λ is the diagonal matrix of the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_{3N-6})$ and V is the matrix of the $3N - 6$ eigenvectors paired to the eigenvalues. The eigenvalues are sorted in descending order with respect to the amount of variance of the original data described by the pairing eigenvectors.

Principal components representations of individual trajectories were generated by projecting the coordinates in the Cartesian space coming from the simulation into the eigenspace defined by the first 3 eigenvectors, as shown in Figures 2 and 3 and Supplementary Material, Figure S7. On the other hand, we quantitatively compared the PCA subspace explored by different mutants and by the wild-type LA5 domain as follows. Using the VMD (88) *CATDCD* utility, we concatenated into a meta-trajectory the last 10 ns of all the trajectories and then recalculated the complete Eigensystem to obtain the projections of the frames of each independent simulation into the meta-trajectory principal components. This approach allows to describe all the different simulations in a common PC space. Then, we quantitatively assessed the effect of mutations on the structure of the LA5 domain by calculating the distance among the subspaces explored by each mutant and the wild-type domain. To do that we used the Mahalanobis distance (62) ($MD_{pp'}$), a metric routinely used in the field of multivariate statistics which, in contrast with the classic Euclidean distance, accounts for the correlations on data and is independent of data transformations. In the specific case of PCA, the Mahalanobis distance between a pair of points p and p' in the PC space is defined as:

$$MD_{pp'} = \sqrt{\sum_{i=0}^{N} \frac{(\mathrm{proj}_{p_i} - \mathrm{proj}_{p'_i})^2}{\lambda_i}}$$

where $proj_{p(p')i}$ are the corresponding projections in the N-dimensional PCA space, and λ_i is the corresponding eigenvalue for the PC_i. The $MD_{pp'}$ normalizes the contributions of all the PCs according to the percentage of variance described by the pairing eigenvector, providing a more realistic assessment of the distance among points in the PCA space.

For obtaining the mean distance among trajectories independently of their compliance (stability) or non-compliance (instability) with the multivariate normal distribution, we set a resampling strategy in which we resampled with replacement a subset of snapshots—e.g. 5 ns for the 10 ns meta-trajectory—from each trajectory, and calculated $MD_{pp'}$ for all possible pairs of points in the N-dimensional PCA space—e.g. 25 eigenvectors describing 95% of the variance. We repeated this step 10^5 times for each pair of simulations and obtained normal distributions for the Mahalanobis distances among points in the trajectories, with rather low-standard deviations. From this comparison, we obtained the mean $MD_{T1,T2}$ among whichever two trajectories. After calculating the distance matrix among trajectories, we performed a clustering—i.e. using a complete-link clustering procedure—of the trajectories according to the PCA subspace explored in each case. All the manipulation of MD data for PCA analysis was performed with a set of *ad hoc* TCL and Perl scripts, alongside with the PCAZIP package (http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html). We compressed all the trajectories using PCAZIP, taking into consideration only the backbone atoms of the LDL-r LA5 domain and retrieving, in each case, the number of eigenvalues and eigenvectors sufficient to describe 95% of the total variance in the system. Using a tool from the PCAZIP package, we extracted all the metrics and data used in the statistical analyses in our study—e.g. eigenvectors, projections etc. The processing of PCA data, and all the resampling, clustering and statistical analyses were done in the R statistical package (92) with a group of *ad hoc* R scripts.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

The authors thankfully acknowledge the resources from the supercomputers *Marenostrum* from the Barcelona Supercomputer Center (BSC), and *Memento* and *Terminus* hosted at the BIFI, Universidad de Zaragoza and the technical expertise and assistance provided by the High Performance Computing groups both at the BSC and BIFI.

Conflict of Interest statement. None declared.

Funding

V.E.A. was funded by Banco Santander Central Hispano, Fundación Carolina and Universidad de Zaragoza and was recipient of a doctoral fellowship awarded by Consejo Superior de Investigaciones Científicas, JAE program. M.O. acknowledge financial support from grants BIO2012–32868 (Ministerio de Economía y Competitividad, Spain), 2014SGR00134 (Grups de Recerca Consolidats, Generalitat de Catalunya, Spain) and PT13/0001/0019 (ISCIII, Spain). J.S. acknowledge financial support from grants BFU2010-16297 (Ministerio de Ciencia e Innovación, Spain), BFU2013-47064-P and BIO2014-57314-REDT (Ministerio de Economía y Competitividad, Spain) and PI078/08 (Gobierno de Aragón, Spain). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding to pay the Open Access publication charges for this article was provided by grant BFU2013-47064-P (Ministerio de Economia Y competitividad).

References

1. Go, G.-W. and Mani, A. (2012) Low-density lipoprotein receptor (LDLR) family orchestrates cholesterol homeostasis. *Yale J. Biol. Med.*, **85**, 19–28.
2. Blacklow, S.C. (2007) Versatility in ligand recognition by LDL receptor family proteins: advances and frontiers. *Curr. Opin. Struct. Biol.*, **17**, 419–426.
3. Nykjaer, A. and Willnow, T.E. (2002) The low-density lipoprotein receptor gene family: a cellular Swiss army knife? *Trends Cell Biol.*, **12**, 273–280.
4. Yamamoto, T., Davis, C.G., Brown, M.S., Schneider, W.J., Casey, M.L., Goldstein, J.L. and Russell, D.W. (1984) The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell*, **39**, 27–38.
5. Guttman, M. and Komives, E.A. (2011) The structure, dynamics, and binding of the LA45 module pair of the low-density lipoprotein receptor suggest an important role for LA4 in ligand release. *Biochemistry*, **50**, 11001–11008.
6. Fisher, C., Beglova, N. and Blacklow, S.C. (2006) Structure of an LDLR-RAP complex reveals a general mode for ligand recognition by lipoprotein receptors. *Mol. Cell*, **22**, 277–283.
7. Beglova, N., Jeon, H., Fisher, C. and Blacklow, S.C. (2004) Cooperation between fixed and low pH-inducible interfaces controls lipoprotein release by the LDL receptor. *Mol. Cell*, **16**, 281–292.
8. Rudenko, G., Henry, L., Henderson, K., Ichtchenko, K., Brown, M.S., Goldstein, J.L. and Deisenhofer, J. (2002) Structure of the LDL receptor extracellular domain at endosomal pH. *Science (New York, NY)*, **298**, 2353–2358.
9. Ren, G., Rudenko, G., Ludtke, S.J., Deisenhofer, J., Chiu, W. and Pownall, H.J. (2010) Model of human low-density lipoprotein and bound receptor based on cryoEM. *Proc. Natl Acad. Sci. USA*, **107**, 1059–1064.
10. Russell, D.W., Brown, M.S. and Goldstein, J.L. (1989) Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. *J. Biol. Chem.*, **264**, 21682–21688.
11. Fisher, C., Abdul-Aziz, D. and Blacklow, S.C. (2004) A two-module region of the low-density lipoprotein receptor sufficient for formation of complexes with apolipoprotein E ligands. *Biochemistry*, **43**, 1037–1044.
12. Boren, J., Lee, I., Zhu, W., Arnold, K., Taylor, S. and Innerarity, T.L. (1998) Identification of the low density lipoprotein receptor-binding site in apolipoprotein B100 and the modulation of its binding activity by the carboxyl terminus in familial defective apo-B100. *J. Clin. Invest.*, **101**, 1084–1093.
13. Arias-Moreno, X., Arolas, J.L., Aviles, F.X., Sancho, J. and Ventura, S. (2008) Scrambled isomers as key intermediates in the oxidative folding of ligand binding module 5 of the low density lipoprotein receptor. *J. Biol. Chem.*, **283**, 13627–13637.
14. Arias-Moreno, X., Velazquez-Campoy, A., Rodríguez, J.C., Pocoví, M. and Sancho, J. (2008) Mechanism of low density lipoprotein (LDL) release in the endosome: implications of the stability and Ca^{2+} affinity of the fifth binding module of the LDL receptor. *J. Biol. Chem.*, **283**, 22670–22679.
15. Abdul-Aziz, D., Fisher, C., Beglova, N. and Blacklow, S.C. (2005) Folding and binding integrity of variants of a prototype ligand-binding module from the LDL receptor possessing multiple alanine substitutions. *Biochemistry*, **44**, 5075–5085.
16. Beglova, N. and Blacklow, S.C. (2005) The LDL receptor: how acid pulls the trigger. *Trends Biochem. Sci.*, **30**, 309–317.
17. Rocha, V.Z., Chacra, A.P.M., Salgado, W., Miname, M., Turolla, L., Gagliardi, A.C.M., Ribeiro, E.E., Rocha, R.P.S., Avila, L.F.R., Pereira, A. *et al.* (2013) Extensive xanthomas and severe subclinical atherosclerosis in homozygous familial hypercholesterolemia. *J. Am. Coll. Cardiol.*, **61**, 2193.
18. Oosterveer, D.M., Versmissen, J., Yazdanpanah, M., Hamza, T.H. and Sijbrands, E.J.G. (2009) Differences in characteristics and risk of cardiovascular disease in familial hypercholesterolemia patients with and without tendon xanthomas: a systematic review and meta-analysis. *Atherosclerosis*, **207**, 311–317.
19. Zambón, D., Quintana, M., Mata, P., Alonso, R., Benavent, J., Cruz-Sánchez, F., Gich, J., Pocoví, M., Civeira, F., Capurro, S. *et al.* (2010) Higher incidence of mild cognitive impairment in familial hypercholesterolemia. *Am. J. Med.*, **123**, 267–274.
20. Guardamagna, O., Restagno, G., Rolfo, E., Pederiva, C., Martini, S., Abello, F., Baracco, V., Pisciotta, L., Pino, E., Calandra, S. *et al.* (2009) The type of LDLR gene mutation predicts cardiovascular risk in children with familial hypercholesterolemia. *J. Pediatr.*, **155**, 199–204.e192.
21. Goldstein, J. and Brown, M. (2009) The LDL receptor. *Arterioscler. Thromb. Vasc. Biol.*, **29**, 431–438.
22. Soutar, A.K. and Naoumova, R.P. (2007) Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nat. Clin. Pract. Cardiovasc. Med.*, **4**, 214–225.
23. Pullinger, C.R., Hennessy, L.K., Chatterton, J.E., Liu, W., Love, J.A., Mendel, C.M., Frost, P.H., Malloy, M.J., Schumaker, V.N. and Kane, J.P. (1995) Familial ligand-defective apolipoprotein B. Identification of a new mutation that decreases LDL receptor binding affinity. *J. Clin. Invest.*, **95**, 1225–1234.
24. Myant, N.B. (1993) Familial defective apolipoprotein B-100: a review, including some comparisons with familial hypercholesterolaemia. *Atherosclerosis*, **104**, 1–18.

25. Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K. and Hobbs, H.H. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.*, **37**, 161–165.
26. Leren, T.P. (2004) Mutations in the PCSK9 gene in Norwegian subjects with autosomal dominant hypercholesterolemia. *Clin. Genet.*, **65**, 419–422.
27. Villéger, L., Abifadel, M., Allard, D., Rabès, J.-P., Thiart, R., Kotze, M.J., Béroud, C., Junien, C., Boileau, C. and Varret, M. (2002) The UMD-LDLR database: additions to the software and 490 new entries to the database. *Hum. Mutat.*, **20**, 81–87.
28. Heath, K.E., Gahan, M., Whittall, R.A. and Humphries, S.E. (2001) Low-density lipoprotein receptor gene (LDLR) worldwide website in familial hypercholesterolaemia: update, new features and mutation analysis. *Atherosclerosis*, **154**, 243–246.
29. Fahed, A.C. and Nemer, G.M. (2011) Familial hypercholesterolemia: the lipids or the genes? *Nutr. Metab.*, **8**, 23.
30. Fass, D., Blacklow, S., Kim, P.S. and Berger, J.M. (1997) Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. *Nature*, **388**, 691–693.
31. Nordestgaard, B.G., Chapman, M.J., Humphries, S.E., Ginsberg, H.N., Masana, L., Descamps, O.S., Wiklund, O., Hegele, R.A., Raal, F.J., Defesche, J.C. *et al.* (2013) 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**, 3478–3490.
32. Lehrman, M.A., Schneider, W.J., Südhof, T.C., Brown, M.S., Goldstein, J.L. and Russell, D.W. (1985) Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science (New York, NY)*, **227**, 140–146.
33. Brown, M.S. and Goldstein, J.L. (1974) Familial hypercholesterolemia: defective binding of lipoproteins to cultured fibroblasts associated with impaired regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity. *Proc. Natl Acad. Sci. USA*, **71**, 788–792.
34. Goldstein, J.L. and Brown, M.S. (1973) Familial hypercholesterolemia: identification of a defect in the regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity associated with overproduction of cholesterol. *Proc. Natl Acad. Sci. USA*, **70**, 2804–2808.
35. UniProt, C. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
36. Peterson, T.A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A. and Kann, M.G. (2010) DMDM: domain mapping of disease mutations. *Bioinformatics (Oxford, England)*, **26**, 2458–2459.
37. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M. and Cooper, D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
38. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
39. Humphries, S.E., Norbury, G., Leigh, S., Hadfield, S.G. and Nair, D. (2008) What is the clinical utility of DNA testing in patients with familial hypercholesterolaemia? *Curr. Opin. Lipidol.*, **19**, 362–368.
40. Varret, M., Abifadel, M., Rabès, J.-P. and Boileau, C. (2008) Genetic heterogeneity of autosomal dominant hypercholesterolemia. *Clin. Genet.*, **73**, 1–13.
41. van Aalst-Cohen, E.S., Jansen, A.C.M., Tanck, M.W.T., Defesche, J.C., Trip, M.D., Lansberg, P.J., Stalenhoef, A.F.H. and Kastelein, J.J.P. (2006) Diagnosing familial hypercholesterolaemia: the relevance of genetic testing. *Eur. Heart J.*, **27**, 2240–2246.
42. Fouchier, S.W., Defesche, J.C., Umans-Eckenhausen, M.W. and Kastelein, J.P. (2001) The molecular basis of familial hypercholesterolemia in The Netherlands. *Hum. Genet.*, **109**, 602–615.
43. Taylor, A., Wang, D., Patel, K., Whittall, R., Wood, G., Farrer, M., Neely, R.D.G., Fairgrieve, S., Nair, D., Barbir, M. *et al.* (2010) Mutation detection rate and spectrum in familial hypercholesterolaemia patients in the UK pilot cascade project. *Clin. Genet.*, **77**, 572–580.
44. Medeiros, A.M., Alves, A.C., Francisco, V., Bourbon, M. and Study, i.o.t.P.F. (2010) Update of the Portuguese Familial Hypercholesterolaemia Study. *Atherosclerosis*, **212**, 553–558.
45. Ned, R.M. and Sijbrands, E.J.G. (2011) Cascade screening for familial hypercholesterolemia (FH). *PLoS Curr.*, **3**, RRN1238. http://currents.plos.org/genomictests/article/cascade-screening-for-familial-70fnx9tmvdav-13/.
46. Zhao, Z. and Michaely, P. (2011) Role of an intramolecular contact on lipoprotein uptake by the LDL receptor. *Biochim. Biophys. Acta*, **1811**, 397–408.
47. Arias-Moreno, X., Cuesta-Lopez, S., Millet, O., Sancho, J. and Velazquez-Campoy, A. (2010) Thermodynamics of protein-cation interaction: Ca(+2) and Mg(+2) binding to the fifth binding module of the LDL receptor. *Proteins*, **78**, 950–961.
48. Huang, S., Henry, L., Ho, Y.K., Pownall, H.J. and Rudenko, G. (2010) Mechanism of LDL binding and release probed by structure-based mutagenesis of the LDL receptor. *J. Lipid Res.*, **51**, 297–308.
49. Cuesta-López, S., Falo, F. and Sancho, J. (2007) Computational diagnosis of protein conformational diseases: short molecular dynamics simulations reveal a fast unfolding of r-LDL mutants that cause familial hypercholesterolemia. *Proteins*, **66**, 87–95.
50. Martínez-Oliván, J., Arias-Moreno, X., Velazquez-Campoy, A., Millet, O. and Sancho, J. (2014) LDL receptor/lipoprotein recognition: endosomal weakening of apo B and apo E binding to the convex face of the LR5 repeat. *FEBS J.*, **281**, 1534–1546.
51. Krivov, G.G., Shapovalov, M.V. and Dunbrack, R.L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.
52. Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)*, **26**, 889–895.
53. Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
54. Laughton, C.A., Orozco, M. and Vranken, W. (2009) COCO: a simple tool to enrich the representation of conformational variability in NMR structures. *Proteins*, **75**, 206–216.
55. Maisuradze, G.G., Liwo, A. and Scheraga, H.A. (2009) Principal component analysis for protein folding dynamics. *J. Mol. Biol.*, **385**, 312–329.
56. Palazoglu, A., Arkun, Y., Erman, B. and Gursoy, A. (2009) Probing protein folding dynamics using multivariate statistical techniques. *Adv. Control Chem. Processes*, **7**, 171–176.
57. David, C.C. and Jacobs, D.J. (2014) Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol. (Clifton, NJ)*, **1084**, 193–226.
58. Berendsen, H.J. and Hayward, S. (2000) Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.*, **10**, 165–169.

59. Van Aalten, D.M.F., De Groot, B.L., Findlay, J.B.C., Berendsen, H.J.C. and Amadei, A. (1997) A comparison of techniques for calculating protein essential dynamics. *J. Comput. Chem.*, **18**, 169–181.
60. Amadei, A., Linssen, A.B. and Berendsen, H.J. (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.
61. Amadei, A., Linssen, A.B., de Groot, B.L., van Aalten, D.M. and Berendsen, H.J. (1996) An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.*, **13**, 615–625.
62. Mahalanobis, P. (1936) On the generalized distance in statistics. *Proc. Natl Inst. Sci. (Calcutta)*, **2**, 49–55.
63. Chiti, F. and Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.
64. Gregersen, N., Bolund, L. and Bross, P. (2005) Protein misfolding, aggregation, and degradation in disease. *Mol. Biotechnol.*, **31**, 141–150.
65. Kopito, R.R. and Ron, D. (2000) Conformational disease. *Nat. Cell Biol.*, **2**, E207–E209.
66. Carrell, R.W. and Lomas, D.A. (1997) Conformational disease. *Lancet*, **350**, 134–138.
67. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
68. Ferrer-Costa, C., Gelpí, J.L., Zamakola, L., Parraga, I., De La Cruz, X. and Orozco, M. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics (Oxford, England)*, **21**, 3176–3178.
69. González-Pérez, A. and López-Bigas, N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
70. Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
71. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
72. Zhao, Z. and Michaely, P. (2008) The epidermal growth factor homology domain of the LDL receptor drives lipoprotein release through an allosteric mechanism involving H190, H562, and H586. *J. Biol. Chem.*, **283**, 26528–26537.
73. Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
74. Reva, B., Antipin, Y. and Sander, C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.*, **8**, R232.
75. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
76. Stone, J.E., Hardy, D.J., Ufimtsev, I.S. and Schulten, K. (2010) GPU-accelerated molecular modeling coming of age. *J. Mol. Graph. Model.*, **29**, 116–125.
77. Bowers, K., Dror, R. and Shaw, D. (2007) Zonal methods for the parallel execution of range-limited N-body simulations. *J. Comput. Phys.*, **221**, 303–329.
78. Fitch, B.G., Rayshubskiy, A., Eleftheriou, M., Ward, T.C., Giampapa, M., Zhestkov, Y., Pitman, M.C., Suits, F., Grossfield, A., Pitera, J. *et al.* (2006), Lecture Notes in Computer Science, Computational Science – ICCS 2006. **3992**, 846–854. Springer, Berlin Heidelberg.
79. Hospital, A., Andrio, P., Fenollosa, C., Cicin-Sain, D., Orozco, M. and Gelpí, J.L. (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics (Oxford, England)*, **28**, 1278–1279.
80. Hospital, A. and Gelpi, J.L. (2013) High-throughput molecular dynamics simulations: toward a dynamic view of macromolecular structure. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 364–377.
81. Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., Repchevsky, D. *et al.* (2010) MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure (London, England : 1993)*, **18**, 1399–1409.
82. Martinez-Olivan, J., Rozado-Aguirre, Z., Arias-Moreno, X., Angarica, V.E., Velazquez-Campoy, A. and Sancho, J. (2014) Low-density lipoprotein receptor is a calcium/magnesium sensor - role of LR4 and LR5 ion interaction kinetics in low-density lipoprotein release in the endosome. *FEBS J.*, **281**, 2638–2658.
83. Martinez-Olivan, J., Arias-Moreno, X., Hurtado-Guerrero, R., Carrodeguas, J.A., Miguel-Romero, L., Marina, A., Bruscolini, P. and Sancho, J. (2015) The closed conformation of the LDL receptor is destabilized by the low Ca(++) concentration but favored by the high Mg(++) concentration in the endosome. *FEBS Lett.*, **589**, 3534–3540.
84. Martinez-Olivan, J., Fraga, H., Arias-Moreno, X., Ventura, S. and Sancho, J. (2015) Intradomain confinement of disulfides in the folding of two consecutive modules of the LDL receptor. *PLoS One*, **10**, e0132141.
85. Garcia-Fandino, R., Bernado, P., Ayuso-Tejedor, S., Sancho, J. and Orozco, M. (2012) Defining the nature of thermal intermediate in 3 state folding proteins: apoflavodoxin, a study case. *PLoS Comput. Biol.*, **8**, e1002647.
86. Lamazares, E., Clemente, I., Bueno, M., Velazquez-Campoy, A. and Sancho, J. (2015) Rational stabilization of complex proteins: a divide and combine approach. *Sci. Rep.*, **5**, 9129.
87. Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
88. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 27–38.
89. Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.
90. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
91. Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **34**, 827–828.
92. R Development Core Team. (2011) R: A Language and Environment for Statistical Computing. the R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.