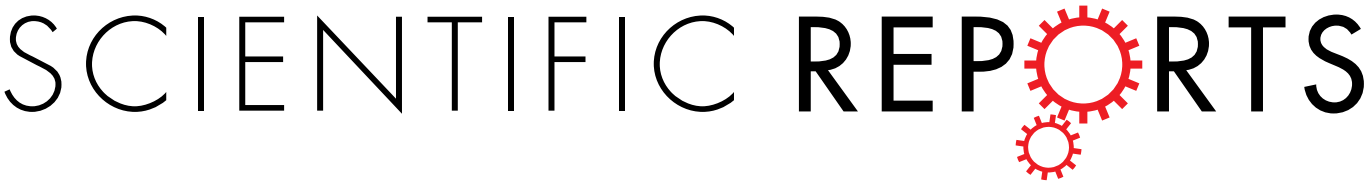

OPEN

In silico analyses of deleterious missense SNPs of human apolipoprotein E3

Received: 21 October 2016
Accepted: 27 March 2017
Published online: 30 May 2017

Allan S. Pires[1,2], William F. Porto[1,2,3], Octavio L. Franco[1,2,4] & Sérgio A. Alencar[1]

ApoE3 is the major chylomicron apolipoprotein, binding in a specific liver peripheral cell receptor, allowing transport and normal catabolism of triglyceride-rich lipoprotein constituents. Point mutations in ApoE3 have been associated with Alzheimer's disease, type III hyperlipoproteinemia, atherosclerosis, telomere shortening and impaired cognitive function. Here, we evaluate the impact of missense SNPs in *APOE* retrieved from dbSNP through 16 computational prediction tools, and further evaluate the structural impact of convergent deleterious changes using 100 ns molecular dynamics simulations. We have found structural changes in four analyzed variants (Pro102Arg, Arg132Ser, Arg176Cys and Trp294Cys), two of them (Pro102Arg and Arg176Cys) being previously associated with human diseases. In all cases, except for Trp294Cys, there was a loss in the number of hydrogen bonds between CT and NT domains that could result in their detachment. In conclusion, data presented here could increase the knowledge of ApoE3 activity and be a starting point for the study of the impact of variations on *APOE* gene.

Apolipoproteins (Apo) compose a family of proteins involved in lipid metabolism, participating in many transport pathways, with major physiological importance. In humans, a large number of apolipoproteins that perform different functions have been described, including ApoA[1,2], ApoB[3,4], ApoC[5], ApoD[6], and ApoE[7,8]. ApoA and ApoD have been described as components of the High Density Lipoprotein (HDL) transport, ApoA being the major component in plasma[2,5], whereas ApoB plays a critical role in the low-density lipoprotein (LDL) transport system[3,4]. Meanwhile, ApoC has been described as a component of very low-density lipoprotein (VLDL)[5]; and ApoE is the major apoliprotein of chylomicrons.

ApoE is capable of binding to a specific liver peripheral cell receptor, allowing transport and normal catabolism of triglyceride-rich lipoprotein constituents[7,8]. It is known that ApoE forms oligomers[9] and, when bound to heparan sulfate proteoglycans (HSPG) and lipids, it adopts an active conformation that allows binding and transport of the low-density lipoprotein receptor (LDLR)[10–12]. Currently, three common isoforms of ApoE are known. These isoforms may be generated by polymorphisms in two different positions within coding regions of the *APOE* gene that lead to amino acid residue changes in positions 130 (site A) and 176 (site B) of the mature ApoE protein: ApoE2 (C130/C176), ApoE3 (C130/R176) and ApoE4 (R130/R176)[9,11,13]. As a result, these differences alter the ApoE function[14,15]. ApoE isoforms have been associated with several human disorders, such as Alzheimer's disease[16,17], type III hyperlipoproteinemia[12,18], atherosclerosis[19], telomere shortening[20], impaired cognitive function[21] and infectious diseases[22,23]. Some of these disorders could be associated with specific isoforms, such as type III hyperlipoproteinemia and Alzheimer's disease, which are associated with ApoE2 and ApoE4[2,16], respectively.

In humans, the most common isoform is ApoE3, characterized as the wild type[12], and this is the unique isoform with a fully elucidated structure, while the other isoforms have only partial structures (e.g. receptor binding domain). Since ApoE3 forms oligomers, some variations (F257A/W264R/V269A/L279Q/V287E) were needed to be inserted in the C-terminus to allow structure elucidation, making a monomeric ApoE3[9]. The ApoE3 structure can be divided into three structural domains: (i) the NT domain comprises the region between residues 1 and 167, (ii) the hinge domain from residues 168 to 205, and (iii) the CT domain from residues 206 to 299[9]. It is known that the CT domain undergoes structural changes when ApoE binds to lipids, leading to activation of

[1]Programa de Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília-DF, Brazil. [2]Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília-DF, Brazil. [3]Porto Reports, Brasília-DF, Brazil. [4]S-Inova Biotech, Pós-graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, MS, Brazil. Correspondence and requests for materials should be addressed to S.A.A. (email: sergiodealencar@gmail.com)

the molecule[9, 12, 15]. Nevertheless, residues 140-160 from the NT domain are important for the interaction with LDL[24, 25]. Furthermore, since the protein is mostly stabilized by hydrogen bond interactions and salt bridges, loss of interactions of this type can cause folding errors or loss of affinity for ligands and they could be involved in disease development[9, 26, 27]. These interactions are very important for the correct folding of CT and Hinge domains[9]. In addition, the interaction between NT and CT exposes hydrophobic residues in CT, increasing lipid affinity[9].

Several studies have shown the effect of point mutations on the functionality of ApoE3. When examining patients with lipoprotein glomerulopathy, Oikawa *et al.* (1991) found that the Arg163Pro point mutation could cause a lower affinity for the LDL receptor (LDLR)[28]. Also, Suehiro *et al.* (1990) demonstrated that the substitution of the same arginine at position 163 of mature protein by a histidine might lead to a lower receptor interaction, increasing the risk for dysbetalipoproteinemia[25, 29]. However, despite the fact that several point mutations present in the coding region of *APOE* have been suggested to be associated with human diseases, the potential impact of missense SNPs described in the dbSNP database has not yet been evaluated.

Currently, computational methods designed to predict the impact of amino acid residue changes in proteins have been widely used in order to assess whether changes are deleterious or not[30]. Among several existing tools, four different groups can be defined based on their methodology: protein-sequence and structure, sequence homology, supervised-learning, and consensus methods[31].

Although there are currently a number of tools used to predict the potential structural and functional impact caused by amino acid changes, these tools are not highly accurate[32]. However, they can still be used as an initial filter of potentially deleterious changes[31]. Then, more refined analysis, such as molecular dynamics simulations, can be used in order to evaluate more precisely the structural impact caused by amino acid changes[31, 33].

The use of molecular dynamics simulations enables the evaluation of structural changes in molecules over a short time window, also allowing observations of changes in physicochemical properties and interactions in simulated environments[34]. However, the use of this method requires high computational power, making it difficult to simulate longer periods. Hence, simulations are limited to just hundreds of nanoseconds. Nevertheless, this method has been widely used to evaluate changes in protein structure caused by point mutations and missense SNPs, such as in the study of α- and β-defensins[35], p53[36], lamin A/C protein[37], guanylin[31], aldosterone synthase[38] and aurora-A kinase[33].

Here, we evaluate the impact of *APOE* missense SNPs from dbSNP by means of a number of computational prediction tools, and further evaluate the structural impact of potentially deleterious changes using molecular dynamics simulations. Our hypothesis is that these variations could cause a significant impact on the protein structure and stability.

Material and Methods

Datasets. The dbSNP database contains SNPs and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants[39]. Using the dbSNP search engine available from the NCBI, only human validated *APOE* SNPs and non-polymorphic single nucleotide variants (SNVs) were filtered. The ApoE3 protein sequence (NCBI Accession: NP_000032.1) was retrieved from the NCBI Protein database (http://www.ncbi.nlm.nih.gov/protein), and the protein structure file of ApoE3 (PDB ID: 2L7B) was obtained from the RCSB Protein Data Bank[9, 40]. The frequency data of missense SNPs found in the *APOE* gene were obtained from the publicly available 1000 Genomes Project (phase I) (http://www.1000genomes.org)[41]. The variant format file (phase 1 release v3.20101123) corresponding to chromosome 19 contained the frequencies of all SNPs identified in the genomes of 1,092 individuals from 14 populations obtained through a combination of low-coverage (2–6x) whole-genome sequence data, targeted deep (50–100x) exome sequencing and dense SNP genotype data. The 14 populations studied were grouped by the predominant component of ancestry into four super-populations: African (AFR) (246 samples), East Asian (ASN) (286 samples), European (EUR) (379 samples) and Ad Mixed American (AMR) (181 samples).

SNP Selection. As rare SNPs occur at very low frequencies (<1%), there is great concern to avoid confounding putative SNPs with sequencing errors common in next-generation sequencing technologies. Therefore, initially we selected from dbSNP only the ones that fit at least one of the following conditions: (i) it has been sequenced in the 1000 Genomes Project; (ii) it has frequency or genotype data (minor alleles observed in at least two chromosomes); and (iii) it has multiple, independent submissions to the refSNP cluster. Then, in order to evaluate the potential functional impact of the obtained *APOE* missense SNPs, we utilized a total of 16 prediction tools, divided into four different methods, as shown below. We filtered all missense SNPs that were classified as deleterious by at least three tools in each of the four groups, and denominated these as convergent deleterious predicted SNPs.

Sequence homology-based methods. The following methods based on sequence homology principles were used to produce missense SNP functional predictions: Sorting Intolerant From Tolerant (SIFT)[42], Provean[43], Mutation Assessor and Panther[44, 45].

Supervised learning methods. Supervised learning algorithms used for missense SNP impact prediction included neural networks (SNAP)[46], support vector machines (MutPred and SuSPect) and random forests (EFIN)[47–49].

Protein sequence and structure-based methods. The following methods either combine information from protein sequence and structure or use protein structural information alone to analyze missense variants: PolyPhen[50], Site Directed Mutator (SDM)[51], Fold-X[52] and PoPMuSiC[53].

Consensus-based methods. In order to obtain a consensus score based on many different SNP impact prediction strategies, the following types of consensus software were used: Condel[54], Meta-SNP[55], PON-P2 and PredictSNP[56,57].

Evolutionary Conservation Analysis. The ConSurf server is a tool for estimating the evolutionary conservation of amino acid positions in a protein molecule based on the phylogenetic relations between homologous sequences[58]. Using the ApoE3 protein sequence (NCBI Accession: NP_000032.1)[40,59], ConSurf, in ConSeq mode, a search was carried out for close homologous sequences using CSI-BLAST (3 iterations and 0.0001 e-value cut-off) against the UNIREF-90 protein database[60,61]. The maximum number of homologs to collect was set as 150, and the minimal and maximal percentage ID between sequences were set as 35 and 95, respectively. The multiple sequence alignment and calculation methods were left as default (MAFFT-L-INS-i and Bayesian). The sequences were then clustered and highly similar sequences removed using CD-HIT[62]. Position-specific conservation scores were computed using the empirical Bayesian algorithm[63].

Signal Peptide Prediction. In order to verify the impact of convergent deleterious SNPs in the signal peptide, Phobius[64] and SignalP 4.0[65] were used for signal peptide topology prediction.

Molecular Modeling. The structural models containing each missense SNP were separately made by means of MODELLER 9.14[66] using the class automodel with default settings. The template used as wild type was the monomeric ApoE3 structure (PDB ID: 2L7B)[59]. One hundred models were generated for each variant. The best models were selected according to DOPE (Discrete Optimized Protein Structure) score, which indicates the most probable structure. The best models were evaluated by PROSA II[67] and PROCHECK[68] softwares. PROSA II evaluates the model quality while PROCHECK evaluates the stereochemical quality of the model through Ramachandran plot. Good quality models were selected by more than 90% of residues in most favoured and additional allowed regions. The visualization of the structures was done in PyMOL (http://www.pymol.org).

Molecular dynamics simulation. The molecular dynamics simulations of the wild type and the four variant structures were performed by GROMACS 4 computational package using the GROMOS96 43A1 force field[69]. Structures are immersed in water cubic boxes with a 12 Å distance between the edge of the box and the protein. The simulations were done under ionic strength conditions (0,2 M NaCl)[70]. The box was filled using the Single Point Charge water model[71]. The dynamics used the wild type and variants three-dimensional models as initial structures. Additional chlorine ions were also inserted into the complexes with positive charges in order to neutralize the system charge. Geometry of water molecules was constrained by using the SETTLE algorithm[72]. Atomic connections were made through LINCS algorithm[73]. Electrostatic corrections were made by Particle Mesh Ewald algorithm[74], with a threshold of 1.4 nm to minimize the computational time. The same cut-off radius was applied for van der Waals interactions. The steepest descent algorithm was applied to minimize system energy for 50,000 steps. After the energy minimization, the temperature (NVT ensemble) and pressure (NPT ensemble) systems were normalized to 300 K and 1 bar, respectively, each per 100 steps. The velocity-rescaling thermostat and the Parrinello-Rahman barostat were used for normalization of temperature and pressure, respectively. Full simulation of the system was made by 100 ns using the leap-frog algorithm as the integrator.

Analyses of molecular dynamics trajectories. Molecular dynamics simulations were analyzed by means of the backbone root mean square deviation (RMSD), radius of gyration (Rg) and solvent accessible surface area (SASA) using the g_rms, g_gyrate and g_sas built in functions of the GROMACS package[69], respectively. The essential dynamics was performed using the g_covar and g_anaeig utilities of the GROMACS package. The number of hydrogen bonds between the NT domain (residues 1–167) and the CT domain (residues 206–299) was analyzed using g_hbond, also from the GROMACS package. In addition, we analyzed the interactions between known regions of the protein previously described by Chen *et al.*[9]. Rg, SASA and the number of hydrogen bonds were plotted as boxplots, because these allow the visualization of the fluctuation and the range in which at least 50% of the data lies.

Results

Distribution and Frequency of *APOE* SNPs. Out of 183 validated *APOE* SNPs, 31 are missense, 21 are synonymous, and two are nonsense variants. There are also 98 intronic, 7 5′ UTR, 6 3′ UTR, 7 downstream, 8 upstream, 1 splice donor and 2 splice acceptor variants. A graphical representation of the distribution of SNPs in the coding and non-coding regions of the gene represented in terms of percentage is shown in Fig. 1. Frequency information was obtained from the 1000 Genomes Project for eight *APOE* missense SNPs (Table S1). All SNPs retrieved from the 1000 Genomes Project are disposed on Table S1. Five of them are rare SNPs with Global Allele Frequency (GAF) values below 1% and occurring only in one of the four populations studied, while the other three variants (Cys130Arg, Arg163Cys and Arg176Cys) have GAFs $\geq$1%. These variants represent ApoE4 (Cys130Arg), ApoE2* (Arg163Cys) and ApoE2 (Arg176Cys).

ApoE3 Convergent Deleterious Predicted SNPs. There are currently a wide variety of computational tools used for predicting the effects of missense SNPs on protein function. In general, depending on the strategy, these tools can be classified into four groups: sequence homology, supervised-learning, protein-sequence and structure, and consensus-based methods. We filtered all missense SNPs that were classified as deleterious by at least three tools in each of the four groups (Table S2). A total of four SNPs (Pro102 Arg, Arg132Ser, Arg176Cys and Trp294Cys), which we previously named as convergent deleterious predicted SNPs[31] were obtained from this filtration (Table 1). Only three SNPs (Thr11Ala, Ala14Thr and Ala18Thr) occur within the signal peptide region, and all remaining SNPs occur within the mature peptide region (Fig. 2A).

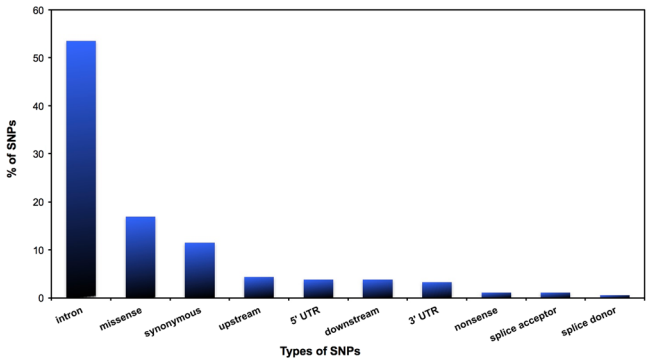

Figure 1. Distribution of SNPs within the APOE gene. The distribution was based on amino acid coding regions (missense, synonymous, nonsense, splice acceptor and splice donor) and on non-coding regions (intronic, upstream and downstream). It can be seen that the majority of the SNPs occur in non-coding regions: 53.6% in introns, 4.4% in upstream regions, 3.8% in 5′ UTR, 3.8% downstream and 3.3% in 3′ UTR. In the coding regions, the majority of the SNPs are missense (16.9%), followed by synonymous (11.5%), nonsense (1.1%), splice acceptor (1.1%) and splice donor (0.6%) variants.

In addition, we analyzed the evolutionary conservation of all missense SNPs within the mature region of ApoE3 using ConSurf[58,75]. ConSurf exploits evolutionary variation in multiple sequence alignments in order to determine the degrees of conservation. The results from this analysis showed that the majority of the variations (66.7%) occur in sites classified as "conserved" (Fig. 2A), including all four convergent deleterious predicted SNPs (Pro102Arg, Arg132Ser, Arg176Cys and Trp294Cys).

The Thr11Ala, Ala14Thr and Ala18Thr Variants Seem not to Alter the Signal Peptide. In order to evaluate the impact of Thr11Ala, Ala14Thr and Ala18Thr in the signal peptide, two prediction servers were used (Phobius and SignalP 4.0). However, none of them indicated any changes in the signal peptide topology.

The impact of variations on protein structure. ApoE3 native structure is characterized by ten α-helices stabilized by hydrogen bonds, salt bridges and hydrophobic interactions (Fig. 2C). The monomeric ApoE3 (PDB ID: 2L7B) was used to construct the variant structures. Since the ApoE3 monomeric structure has some modifications in the C-terminal, we modeled the native structure by the substitution of respective residues in the C-terminus. Table S3 summarizes the validation assessments. We performed molecular dynamics simulations to evaluate which probable structural changes occur within each modelled ApoE3 structure. The best model for each variant was simulated for 100 ns. The analysis of RMSD was carried out to measure differences in movement between native and variant backbones. The RMSD analysis showed that the native structure had little variation during the simulation time, ranging from 3 to 4 Å (Fig. 3A). Despite that, all analyzed variants presented a higher variation in the backbone of the protein ranging from 3 to 6 Å in Pro102Arg and Arg132Ser and from 3 to 5 Å in Arg176Cys and Trp294 simulations (Fig. 3A).

In contrast, analysis of the radius of gyration showed wide differences between variant and wild structures, with an increase for all variants (Fig. 3B). The protein flexibility was also analyzed, by means of essential dynamics, showing that all the variants had a gain in flexibility (Figure S1). Therefore, solvent accessible surface area and radius of gyration were measured in order to evaluate the maintenance of protein packing. The solvent accessible surface area analysis of the variant structures showed little difference between the wild type structure and Arg132Ser and Arg176Cys variants, with little or no increase (Fig. 3C). However, Pro102Arg and Trp294Cys showed a higher increase on this property (Fig. 3C).

Structural changes in CT may reduce the ApoE3 affinity to lipids[59]. Since NT stabilizes CT, we verified whether some of the convergent deleterious SNPs could affect the number of hydrogen bonds made between CT and NT amino acid residues. There were differences between wild type and variant structures in all cases. While the Arg132Ser and Trp294Cys variants showed a decrease in the number of hydrogen bonds in comparison to the wild type structure (Fig. 3D), the Pro102Arg variant exhibited an increase (Fig. 3D). Moreover, Arg176Cys showed a little increase in the number of hydrogen bonds in comparison to the wild type structure, however, almost the same behavior as the wild type (Fig. 3D). Furthermore, we analyzed differences in the number of interactions between known structural regions in native and variant structures over time. From this, we measured the variant effects on known interactions of native structure. In analyzes with NT and CT domains, almost all variants presented differences when compared to the native structure, with a decrease in Arg132Ser and Trp294Cys variants, an increase in Pro102Arg and the same number of interactions in Arg176Cys (Fig. 3D). However, only Trp294Cys presented a loss of hydrogen bonds between known regions in ApoE3 (Figure S2 and Table S4). Meanwhile, the other three variants presented a great increase in these interactions. However, Pro102Ser presented the greatest impact on the number of hydrogen bonds between the structural domains, with an average gain of about 17 hydrogen bonds in relation to the native structure.

SNP rs [#]	Amino Acid Change[a]	ValidationMethod[b]	Sequence-Based[c]				SLM-Based[c]				Consensus-Based[c]				Structure-Based[c]			
			SIFT	Provean	Mutation Assessor	Panther	MutPred	EFIN	SNAP	SuSPect	Condel	MetaSNP	PON-P2	Predict SNP	PolyPhen	SDM	Fold-X	PoPMuSiC
rs11083750:C > A	Pro102Arg	Cluster	D	D	D	U	N	D	D	D	D	D	P	D	D	N	DT	DT
rs11542041:C > A	Arg132Ser	1000 G	D	D	D	D	D	D	D	D	N	D	P	D	D	D	DT	DT
rs7412:C > T	Arg176Cys	1000 G, cluster, freq.	D	D	D	D	D	D	D	D	D	D	N	D	D	N	DT	DT
rs557715042:G > T	Trp294Cys	1000 G, freq.	D	D	D	U	D	D	D	N	D	N	P	D	D	D	DT	DT

Table 1. Results of *APOE* convergent deleterious predicted SNPs analyzed by 16 prediction tools classified in four different groups. [a]APOE amino acid positions is relative to GenBank Accession number NP_000032.1. [b]1000G: SNP has been sequenced in the 1000 Genomes Project; freq.: Validated by frequency or genotype data: minor alleles observed in at least two chromosomes; cluster: Validated by multiple, independent submissions to the refSNP cluster. [c]N: Neutral; D: Deleterious; ST: Stabilizing; DT: Destabilizing; P: Pathogenic; U: Unknown.

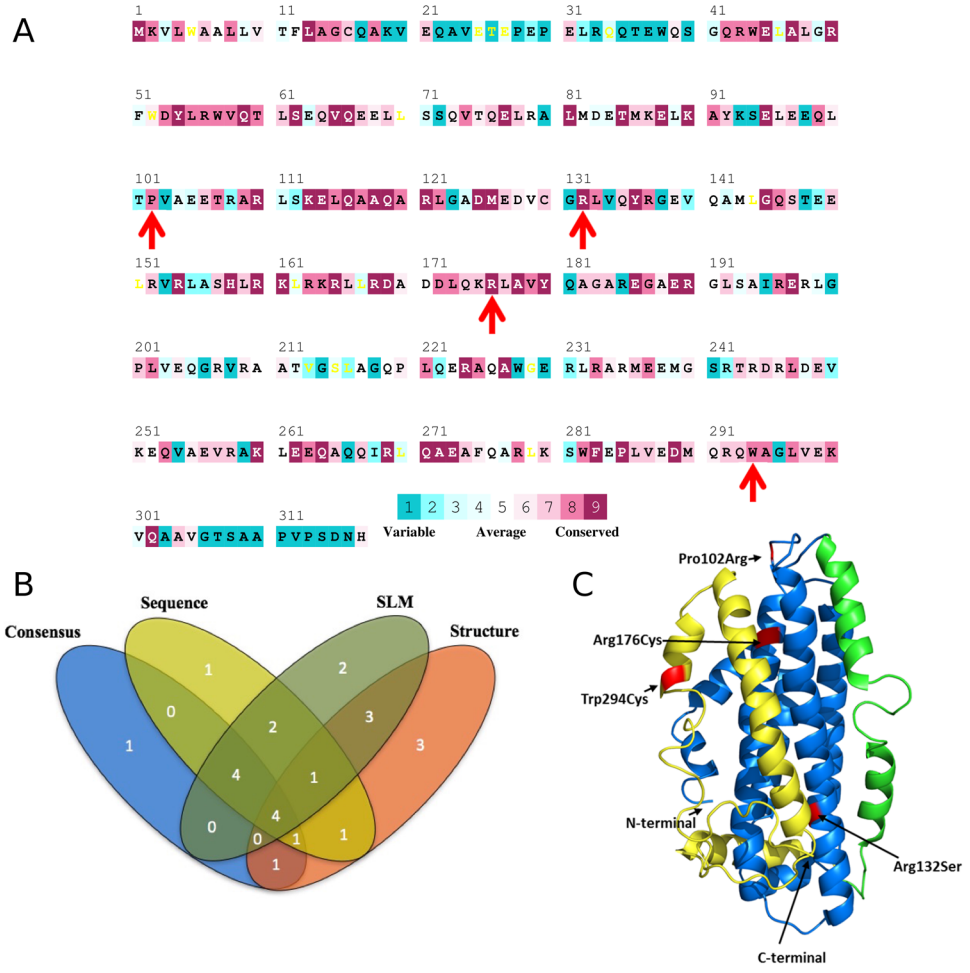

Figure 2. Missense SNPs identified in the APOE gene and Structural domains of native ApoE3. Conservation pattern of amino acid residues within the mature peptide region of ApoE3 obtained from multiple sequence alignment using ConSurf. Color intensity increases with degree of conservation. The amino acids are coloured based on their conservation grades and conservation levels. A grade of 1 indicates rapidly evolving (variable) sites, which are colour-coded in turquoise; 5 indicates sites that are evolving at an average rate, which are coloured white; and 9 indicates slowly evolving (evolutionarily conserved) sites, which are colour-coded in maroon. The four convergent deleterious predicted SNPs are marked below the peptide sequence as red arrows (**A**). Venn diagram showing the relationships between missense SNPs predicted as deleterious by the four different groups (sequence homology, supervised-learning (SLM), protein-sequence and structure, and consensus-based methods). A total of four convergent deleterious predicted SNPs (classified as deleterious by at least three tools in each of the four different groups) were obtained (**B**). Structural domains of native ApoE3. In blue is represented the NT domain, CT domain is represented in yellow and hinge region is showed in green. In red are highlighted the different variations analyzed in this work, identified by arrows (**C**).

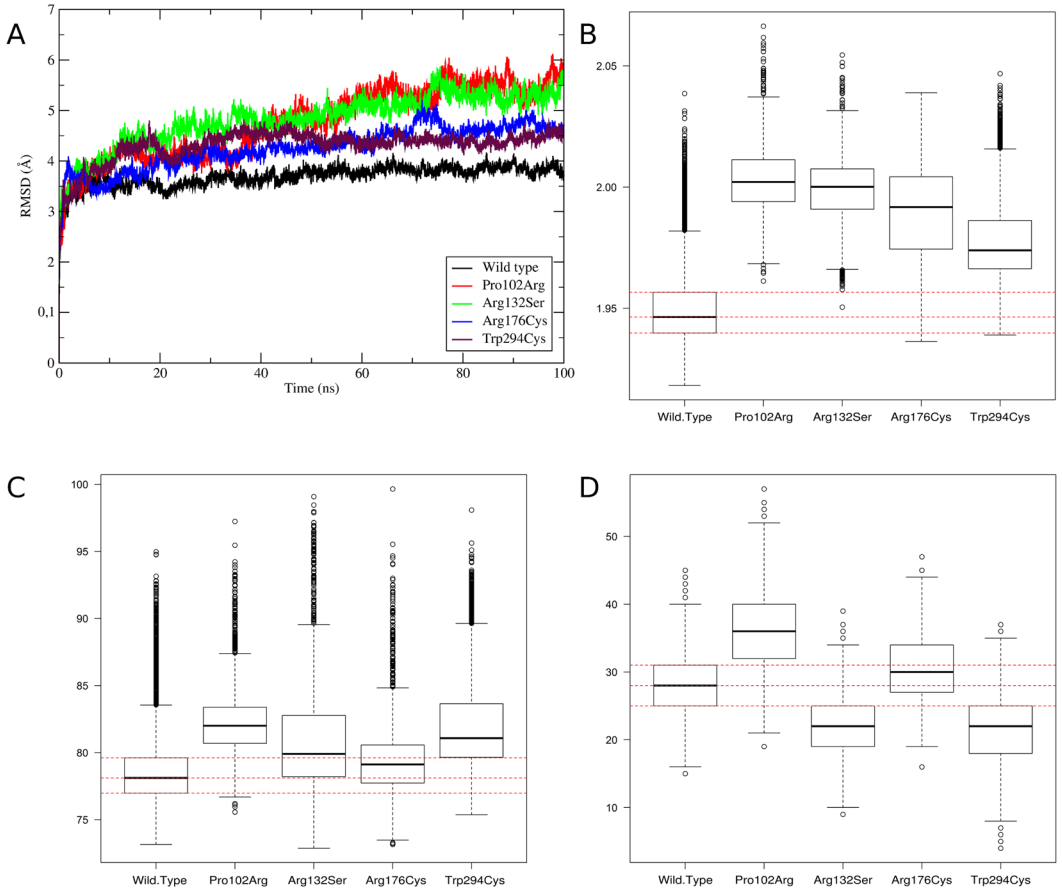

Figure 3. ApoE3 native and variants trajectories analyses. In Backbone RMSD variation the variants are identified in the plots by different colors (**A**). Radius of gyration (**B**), solvent accessible surface area (**C**) and number of hydrogen bonds (**D**) are plotted in boxplots. On backbone RMSD (**A**) the variants are identified in the plots by colors: Native structure (black), Pro102Arg (red), Arg132Ser (green), Arg176Cys (blue) and Trp296Cys (turquoise). Only the number of hydrogen bonds between NT and CT domains was computed (**D**). Dotted red lines on solvent accessible surface area, radius of gyration and number of hydrogen bonds plots indicate the reference values of wild type. The solvent accessible surface and radius of gyration values are in nm^2 and to RMSD values in Å.

Discussion

ApoE3 presents an helical structure stabilized by hydrogen bonds and salt bridges[59]. This characteristic confers protein plasticity and capacity of large conformational changes, important for the activity performed by this protein. Here, we used molecular dynamics simulations to assess conformational changes caused by the presence of missense SNPs that lead to amino acid residue changes in the coded protein. We were able to simulate a protein of 299 amino acid residues for 100 ns. For short peptides, it is not difficult to reach this simulation time[31], but for proteins greater than 200 amino acids it is common to simulate for less than 10 ns[36–38], with few exceptions being simulated for more than 100 ns[33].

All of the four variants analyzed here are present in conserved regions of the protein (Fig. 2B). Therefore, the implementation of 16 prediction tools to pre-filter potentially damaging SNPs present in the *APOE* gene could in fact lead to the discovery of variations that have an impact on protein structure and, consequently, on its function. Furthermore, the use of a consensus of different types of tools (e.g sequence homology-based, supervised learning method and protein sequence and structure-based) to screen potentially damaging SNPs increases their prediction accuracy. Out of the four variants, Pro102Arg presented an increase in all analyses compared to the native protein (Fig. 3D). Interestingly, despite a gain in the number of hydrogen bonds between both CT and NT, as well as between known structural domains, this variant presents the largest differences relative to the wild type structure (Fig. 3).

However, this variant has not been associated with any diseases reported in the literature yet. It is known that ApoE4 is associated with hyperlipidemia[2], nevertheless, the double mutant (Cys112Arg/Pro102Arg) has not been described as having this association[76,77]. Despite the compensatory effect of Pro102Arg on ApoE4, in ApoE3 it could be deleterious due to the gain in radius of gyration, surface and hydrogen bonds.

On the other hand, the Trp294Cys and Arg132Ser variants presented loss in hydrogen bonds between CT and NT domains (Fig. 3D). This occurs due to the loss of a hydrophobic amino acid in the CT domain. Besides

that, the substitution of Trp294 could interfere with lipid interaction mediated by the CT domain, causing loss of affinity[59]. This step of interaction with lipid was previously associated with activation of the protein, starting the essential structural changes that expose the LDLR binding region in the NT domain[59]. Moreover, previously, single point changes in CT were used to inhibit the oligomerization of ApoE3[59]. Therefore, it is possible that missense SNPs present in this region could also interpose the normal behavior of the protein. On the other hand, Arg132Ser is important in interdomain interaction, performing two hydrogen bonds with CT domain residues (Gln235 and Glu238)[59]. Then, loss of hydrogen bonds caused by Arg132Ser could generate the separation of the CT and NT domains, exposing the LDLR interaction domain without the activation by lipids[59].

Finally, the Arg176Cys variant presented a more similar behavior compared to the native protein (Fig. 3). Despite this, given the large variation in the radius of gyration analysis of the Arg176Ser variant and the increase of the RMSD, it is possible that this variant generates an opening and closing movement of the Arg176Ser variant, which causes the highest variation on radius of gyration. The Arg176Cys variant characterizes the E2 isoform, which is associated with diseases such as hyperlipoproteinemia III[78,79] and atherosclerosis[78]. Our analysis showed that this variant could result in a change in affinity between ApoE and LDLR, generating the clinical condition[13,24,59,80]. The Arg176Cys variant did not show great differences in number of hydrogen bonds (Fig. 3D) or solvent accessible surface analyses (Fig. 3C). Furthermore, this variation has a GAF $\geq$1%, being the most common variation in this study.

Conclusions

Although many variations have been identified in the *APOE* gene, the potential structural and functional impact of many of them have not been analyzed yet. However, the four analyzed variants could lead the protein to lose affinity with lipids. The loss of hydrogen bonds between NT and CT domains viewed in variants may be an important factor for research into association between diseases and ApoE variations. Furthermore, the similarity in ApoE2 and other variations could be significant to analyses of impact of these variations and their association with diseases. In conclusion, data presented here could increase the knowledge of ApoE3 activity and be a starting point for the study of impact of variations on the *APOE* gene.

References

1. Narayanaswami, V. & Ryan, R. O. Molecular basis of exchangeable apolipoprotein function. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1483**, 15–36, doi:10.1016/S1388-1981(99)00176-6 (2000).
2. Breslow, J. L. *et al.* Isolation and characterization of cDNA clones for human apolipoprotein A-I. *Proc. Natl. Acad. Sci. USA* **79**, 6861–6865, doi:10.1073/pnas.79.22.6861 (1982).
3. Lusis, A. J. *et al.* Cloning and expression of apolipoprotein B, the major protein of low and very low density lipoproteins. *Proc. Natl. Acad. Sci. USA* **82**, 4597–4601, doi:10.1073/pnas.82.14.4597 (1985).
4. Law, S. W. *et al.* Human apolipoprotein B-100: cloning, analysis of liver mRNA, and assignment of the gene to chromosome 2. *Proc. Natl. Acad. Sci. USA* **82**, 8340–8344, doi:10.1073/pnas.82.24.8340 (1985).
5. Vaith, P., Assmann, G. & Uhlenbruck, G. Characterization of the oligosaccharide side chain of apolipoprotein C-III from human plasma very low density lipoproteins. *Biochim. Biophys. Acta* **541**, 234–240, doi:10.1016/0304-4165(78)90396-3 (1978).
6. Rassart, E. *et al.* Apolipoprotein D. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1482**, 185–198, doi:10.1016/S0167-4838(00)00162-X (2000).
7. Utermann, G., Weber, W. & Beisiegel, U. Different mobility in SDS-polyacrylamide gel electrophoresis of Apolipoprotein E from phenotypes Apo E-N and Apo E-D. *FEBS Lett.* **101**, 21–26, doi:10.1016/0014-5793(79)81286-7 (1979).
8. Utermann, G., Pruin, N. & Steinmetz, A. Polymorphism of apolipoprotein E. III. Effect of a single polymorphic gene locus on plasma lipid levels in man. *Clin. Genet.* **15**, 63–72, doi:10.1111/j.1399-0004.1979.tb02028.x (1979).
9. Chen, J., Li, Q. & Wang, J. Topology of human apolipoprotein E3 uniquely regulates its diverse biological functions. *Proc. Natl. Acad. Sci. USA* **108**, 14813–14818, doi:10.1073/pnas.1106420108 (2011).
10. Hatters, D. M., Peters-Libeu, C. A. & Weisgraber, K. H. Apolipoprotein E structure: insights into function. *Trends Biochem. Sci.* **31**, 445–454, doi:10.1016/j.tibs.2006.06.008 (2006).
11. Weisgraber, K. H. & Apolipoprotein, E. structure-function relationships. *Adv. Protein Chem.* **41**, 853–72 (1994).
12. Mahley, R. W., Weisgraber, K. H. & Huang, Y. Apolipoprotein E: structure determines function, from atherosclerosis to Alzheimer's disease to AIDS. *J. Lipid Res.* **50**(Suppl), S183–S188, doi:10.1194/jlr.R800069-JLR200 (2009).
13. Weisgraber, K. H., Rall, S. C. & Mahley, R. W. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J. Biol. Chem.* **256**, 9077–9083 (1981).
14. Mahley, R. W., Huang, Y. & Rall, S. C. Jr. Pathogenesis of type III hyperlipoproteinemia (dysbetalipoproteinemia). Questions, quandaries, and paradoxes. *J Lipid Res* **40**, 1933–1949 (1999).
15. Zuo, L. *et al.* Variation at APOE and STH loci and Alzheimer's disease. *Behav. Brain Funct.* **2**, 13, doi:10.1186/1744-9081-2-13 (2006).
16. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923, doi:10.1126/science.8346443 (1993).
17. Wolk, Da & Dickerson, B. C. Apolipoprotein E (APOE) genotype has dissociable effects on memory and attentional-executive network function in Alzheimer's disease. *Proc. Natl. Acad. Sci. USA* **107**, 10256–10261, doi:10.1073/pnas.1001412107 (2010).
18. Marais, A. D., Solomon, G. A. E. & Blom, D. J. Dysbetalipoproteinaemia: A mixed hyperlipidaemia of remnant lipoproteins due to mutations in apolipoprotein E. *Crit. Rev. Clin. Lab. Sci.* **51**, 46–62, doi:10.3109/10408363.2013.870526 (2014).
19. McNeill, E., Channon, K. M. & Greaves, D. R. Inflammatory cell recruitment in cardiovascular disease: murine models and potential clinical applications. *Clin. Sci. (Lond).* **118**, 641–55, doi:10.1042/CS20090488 (2010).
20. Jacobs, E. G. *et al.* Accelerated Cell Aging in Female APOE-?? 4 Carriers: Implications for Hormone Therapy Use. *PLoS One* **8**, (2013).
21. Deary, I. J. *et al.* Cognitive change and the APOE epsilon 4 allele. *Nature* **418**, 932–932, doi:10.1038/418932a (2002).
22. Burt, T. D. *et al.* Apolipoprotein (apo) E4 enhances HIV-1 cell entry *in vitro*, and the APOE epsilon4/epsilon4 genotype accelerates HIV disease progression. *Proc. Natl. Acad. Sci. USA* **105**, 8718–8723, doi:10.1073/pnas.0803526105 (2008).
23. de Bont, N. *et al.* Apolipoprotein E knock-out mice are highly susceptible to endotoxemia and Klebsiella pneumoniae infection. *J. Lipid Res.* **40**, 680–685 (1999).
24. Innerarity, T. L., Friedlander, E. J., Rall, S. C., Weisgraber, K. H. & Mahley, R. W. The receptor-binding domain of human apolipoprotein E. Binding of apolipoprotein E fragments. *J. Biol. Chem.* **258**, 12341–12347 (1983).
25. Suehiro, T., Yoshida, K. & Yamano, T. of a NewVariant of Apolipoprotein E (apo E-Kochi). **29**, 587–594 (1990).
26. Weisgraber, K. H. & Mahley, R. W. Human apolipoprotein E: the Alzheimer's disease connection. *FASEB J.* **10**, 1485–94 (1996).

27. Mahley, R. W. & Huang, Y. Apolipoprotein (apo) E4 and Alzheimer's disease: Unique conformational and biophysical properties of apoE4 can modulate neuropathology. *Acta Neurol. Scand.* **114**, 8–14, doi:10.1111/j.1600-0404.2006.00679.x (2006).
28. Oikawa, S. *et al*. Abnormal lipoprotein and apolipoprotein pattern in lipoprotein glomerulopathy. *Am. J. Kidney Dis.* **18**, 553–558, doi:10.1016/S0272-6386(12)80649-4 (1991).
29. Ishigaki, Y. *et al*. Virus-mediated transduction of apolipoprotein E (ApoE)-Sendai develops lipoprotein glomerulopathy in ApoE-deficient mice. *J. Biol. Chem.* **275**, 31269–31273, doi:10.1074/jbc.M005906200 (2000).
30. Zhang, Z., Miteva, M. A., Wang, L. & Alexov, E. Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med*. **2012**, (2012).
31. Porto, W. F., Franco, O. L. & Alencar, S. a. Computational analyses and prediction of guanylin deleterious SNPs. *Peptides* 1–11, 10.1016/j.peptides.2015.04.013 (2015).
32. Rodrigues, C., Santos-Silva, A., Costa, E. & Bronze-da-Rocha, E. Performance of *In Silico* Tools for the Evaluation of UGT1A1 Missense Variants. *Hum. Mutat.* **36**, 1215–1225, doi:10.1002/humu.22903 (2015).
33. Kumar, A. & Purohit, R. Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs. *PLoS Comput. Biol.* **10**, (2014).
34. Hospital, A., Goñi, J. R., Orozco, M. & Gelpi, J. Molecular dynamics simulations: Advances and applications. *Adv. Appl. Bioinforma. Chem.* **8**, 37–47, doi:10.2147/AABC.S70333 (2015).
35. Porto, W. F., Nolasco, D. O., Pires, Á. S., Pereira, R. W. & Octávio, L. Prediction of the Impact of Coding Missense and Nonsense Single Nucleotide Polymorphisms on HD5 and HBD1 Antibacterial Activity against *Escherichia coli*. *Biopolym. Pept. Sci*. 1–36 (2016).
36. Chitrala, K. N. & Yeguvapalli, S. Computational screening and molecular dynamic simulation of breast cancer associated deleterious non-synonymous single nucleotide polymorphisms in TP53 gene. *PLoS One* **9**, (2014).
37. Rajendran, V., Purohit, R. & Sethumadhavan, R. *In silico* investigation of molecular mechanism of laminopathy caused by a point mutation (R482W) in lamin A/C protein. *Amino Acids* **43**, 603–615, doi:10.1007/s00726-011-1108-7 (2012).
38. Jia, M. *et al*. Computational analysis of functional single nucleotide polymorphisms associated with the CYP11B2 gene. *PLoS One* **9**, (2014).
39. Sherry, S. T. *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311, doi:10.1093/nar/29.1.308 (2001).
40. Bernstein, F. C. *et al*. The protein data bank: A computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **185**, 584–591, doi:10.1016/0003-9861(78)90204-7 (1978).
41. 1000 Genomes Project Consortium, T. 1000 G. P. *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, doi:10.1038/nature11632 (2012).
42. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081, doi:10.1038/nprot.2009.86 (2009).
43. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7**, e46688, doi:10.1371/journal.pone.0046688 (2012).
44. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118–e118, doi:10.1093/nar/gkr407 (2011).
45. Mi, H. *et al*. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*. **33** (2005).
46. Bromberg, Y., Yachdav, G. & Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**, 2397–2398, doi:10.1093/bioinformatics/btn435 (2008).
47. Zeng, S., Yang, J., Chung, B. H.-Y., Lau, Y. L. & Yang, W. EFIN: predicting the functional impact of nonsynonymous single nucleotide polymorphisms in human genome. *BMC Genomics* **15**, 455, doi:10.1186/1471-2164-15-455 (2014).
48. Flicek, P. *et al*. Ensembl 2014. *Nucleic Acids Res*. **42**, (2014).
49. Li, B. *et al*. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750, doi:10.1093/bioinformatics/btp528 (2009).
50. Adzhubei, I. A. *et al*. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249, doi:10.1038/nmeth0410-248 (2010).
51. Worth, C. L., Preissner, R. & Blundell, T. L. SDM - A server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* **39**, W215–W222, doi:10.1093/nar/gkr363 (2011).
52. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387, doi:10.1016/S0022-2836(02)00442-4 (2002).
53. Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* **12**, 151, doi:10.1186/1471-2105-12-151 (2011).
54. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88**, 440–449, doi:10.1016/j.ajhg.2011.03.004 (2011).
55. Capriotti, E., Altman, R. B. & Bromberg, Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics* **14**(Suppl 3), S2, doi:10.1186/1471-2164-14-S3-S2 (2013).
56. Bendl, J. *et al*. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.* **10**, e1003440, doi:10.1371/journal.pcbi.1003440 (2014).
57. Niroula, A., Urolagin, S. & Vihinen, M. PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One* **10**, (2015).
58. Celniker, G. *et al*. ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry* **53**, 199–206, doi:10.1002/ijch.v53.3/4 (2013).
59. Chen, J. *et al*. Apolipoprotein E and Alzheimer's Disease A Role in Amyloid Catabolism. *roc. Natl. Acad. Sci. USA* **256**, 9077–9083 (2010).
60. Angermüller, C., Biegert, A. & Söding, J. Discriminative modelling of context-specific amino acid substitution probabilities. *Bioinformatics* **28**, 3240–3247, doi:10.1093/bioinformatics/bts622 (2012).
61. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288, doi:10.1093/bioinformatics/btm098 (2007).
62. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659, doi:10.1093/bioinformatics/btl158 (2006).
63. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791, doi:10.1093/molbev/msh194 (2004).
64. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432, doi:10.1093/nar/gkm256 (2007).
65. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–6, doi:10.1038/nmeth.1701 (2011).
66. Fiser, A. & Šali, A. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods Enzymol* **374**, 461–491, doi:10.1016/S0076-6879(03)74020-8 (2003).
67. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **35**, W407–W410, doi:10.1093/nar/gkm290 (2007).

68. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291, doi:10.1107/S0021889892009944 (1993).
69. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447, doi:10.1021/ct700301q (2008).
70. Ibragimova, G. T. & Wade, R. C. Importance of explicit salt ions for protein stability in molecular dynamics simulation. *Biophysical Journal.* **74**, 2906–2911, doi:10.1016/S0006-3495(98)77997-4 (1998).
71. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. Interaction Models For Water In Relation To Protein Hydration. *Intermol. Forces* **31**, 331–338, doi:10.1007/978-94-015-7658-1 (1981).
72. Miyamoto, S. & Kollman, P. A. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
73. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472, doi:10.1002/(ISSN)1096-987X (1997).
74. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092, doi:10.1063/1.464397 (1993).
75. Berezin, C. *et al*. ConSeq: The identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **20**, 1322–1324, doi:10.1093/bioinformatics/bth070 (2004).
76. Ordovas, J. M., Litwack-Klein, L., Wilson, P. W., Schaefer, M. M. & Schaefer, E. J. Apolipoprotein E isoform phenotyping methodology and population frequency with identification of apoE1 and apoE5 isoforms. *J. Lipid Res.* **28**, 371–380 (1987).
77. Wardell, M. R., Rall, S. C., Schaefer, E. J., Kane, J. P. & Weisgraber, K. H. *Two apolipoprotein E5 variants illustrate the importance of the position of additional positive charge on receptor-binding activity. Journal of lipid research* **32**, 521–528 (1991).
78. Sullivan, P. M., Mezdour, H., Quarfordt, S. H. & Maeda, N. Type III hyperlipoproteinemia and spontaneous atherosclerosis in mice resulting from gene replacement of mouse Apoe with human APOE*2. *J. Clin. Invest.* **102**, 130–135, doi:10.1172/JCI2673 (1998).
79. Rall, S. C. *et al*. Type III hyperlipoproteinemia associated with apolipoprotein E phenotype E3/3. Structure and genetics of an apolipoprotein E3 variant. *J. Clin. Invest.* **83**, 1095–1101, doi:10.1172/JCI113988 (1989).
80. Mahley, R. W. & Rall, S. C. Apolipoprotein E: far more than a lipid transport protein. *Annu. Rev. Genomics Hum. Genet.* **1**, 507–37, doi:10.1146/annurev.genom.1.1.507 (2000).

Acknowledgements

This work was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior); FAPDF (Fundação de Amparo à Pesquisa do Distrito Federal); FUNDECT (Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul) and UCB (Universidade Católica de Brasília).

Author Contributions

Conceived and designed the experiments: W.P. and S.A. Analyzed the data: A.P., W.P. and S.A. Wrote the main manuscript text: A.P., W.P., O.L.F. and S.A. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-01737-w

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.