Evolutionary and Structural Constraints Influencing Apolipoprotein A-I Amyloid Behaviour

Gisonno RA[a], #, Masson T[b,#], Ramella N[a],, Barrera EE[c], Romanowski V[b], Tricerri MA[a]

[a] Instituto de Investigaciones Bioquímicas de La Plata (INIBIOLP, CONICET-UNLP), Facultad de Ciencias Médicas, Universidad Nacional de La Plata, La Plata, Argentina [b] Instituto de Biotecnología y Biología Molecular (IBBM, CONICET-UNLP), Facultad de Ciencias Exactas, Universidad Nacional de La Plata), La Plata, Argentina [c] Group of Biomolecular Simulations, Institut Pasteur de Montevideo, Montevideo, Uruguay

[#] Co-first authors

Highlights

- Aggregation prone region 1 (APR1), comprising residues 14-19, is consistently conserved during the evolutionary history of Apolipoprotein A-I.
- APR1 contributes to thermal stability of the -helix bundle in the full-length Apolipoprotein A-I model.
- Amyloid variants introduce a destabilizing effect on the monomer structure of Apolipoprotein A-I, in contrast to HDL-deficiency and gnomAD variants, which are nearly neutral.
- During molecular dynamics simulations, G26R mutant lead to the partial unfolding of -helix bundle and exposure of APR1.

Abstract

Apolipoprotein A-I (apoA-I) has a key function in the reverse cholesterol transport mediated by the high density lipoprotein (HDL) particles. However, aggregation of apoA-I single point mutants can lead to hereditary amyloid pathology. Although several studies have tackled the biophysical and structural impacts introduced by these mutations, there is little information addressing the relationship between the evolutionary and structural features that contribute to the amyloid behaviour of apoA-I. We combined evolutionary studies, in silico saturation mutagenesis and molecular dynamics (MD) simulations to provide a comprehensive analysis of the conservation and pathogenic role of the aggregation prone regions (APRs) present in apoA-I. ApoA-I sequences analysis demonstrated the pervasive conservation of an APR, designated APR1, within the N-terminal -helix bundle. Moreover, stability analysis carried out with the FoldX engine showed that this region contributes to the marginal stability of apoA-I. Structural properties of the full-length apoA-I model suggest that aggregation is avoided by placing APRs into highly packed and rigid portions of its structure. Compared to HDL-deficiency or natural silent variants extracted from the gnomAD database, the thermodynamic and pathogenic impact of apoA-I point mutations associated with amyloid pathologies were found to show a higher destabilizing effect. MD simulations of the amyloid variant G26R evidenced the partial unfolding of the -helix bundle and the occurrence of -strand secondary elements at the C-terminus of apoA-I. Our findings highlight APR1 as a relevant component for apoA-I structural integrity and emphasize a destabilizing effect of amyloid variants that leads to the exposure of APRs. This information contributes to our understanding of how apoA-I, with its high degree of structural flexibility, maintains a delicate equilibrium between its lipid-binding function and intrinsic tendency to form amyloid aggregates. In addition, our stability measurements could be used as a proxy to interpret the structural impact of new mutations affecting apoA-I.

Introduction

Apolipoprotein A-I (apoA-I) is the most abundant protein component of high-density lipoproteins (HDL) and is responsible for the reverse cholesterol transport from extracellular tissues back to the liver (Lund-Katz and Phillips (2010); Rader et al. (2009)), which has been associated with a protective function against cardiac disease and atherosclerosis (Navab et al. (2009); Rosenson et al. (2015)). The scaffolding functions of apoA-I in the HDL particle and its multiple protein-protein interactions, mainly with the lecithin:cholesterol acyltransferase and the ATP-binding cassette A1 transporter (Chroni et al. (2003); Manthei et al. (2020)), forces it to maintain a dynamical and flexible conformation (Gursky and Atkinson (1996)).

In contrast to these physiological functions, several point mutations affecting apoA-I have been associated with hereditary amyloid pathology (Sipe et al. (2016)). These mutations are mainly distributed into two "hot spots," located at the N-terminal region and the C-terminal portion of the protein, each one with a typical clinical manifestation (Das and Gursky (2015)). Mutations that occur at the N-terminal region (residues 26–100) are characterized by amyloid deposits in the liver and kidney (Mucchiano et al. (2001); Obici et al. (2006)), while those located at a short C-terminal domain (residues 170–178) are mainly associated with heart, larynx and skin deposits (Gaglione et al. (2018)). In non-hereditary amyloidosis, full-length apoA-I is deposited in atherosclerotic plaques as fibrils or the senile forms of amyloid. This process has been associated with aging, but it has also been described in chronic pathologies such as Alzheimer's disease and type 2 diabetes mellitus ((**Westermark_1995?**)).

Amyloid behaviour of apoA-I N-terminal fragment has been attributed to the presence of aggregation-prone regions (APRs) in its sequence and, specifically, to an APR located at the N-terminus (Obici et al. (2006)). It has been hypothesized that amyloidogenic mutations or post-translational modifications could promote aggregation through the destabilization of the partially disorganized structure of apoA-I -described as a molten globular state- followed by the exposure of APRs. In this sense, most studies addressing the effect of amyloid variants have focused on the biophysical and physiological consequences of specific mutants. However, our understanding of the relationship between apoA-I sequence determinants and its aggregation process remains unclear.

In this study, through an extensive evolutionary analysis we characterized the conservation of aggregating regions in a broad dataset of vertebrates apoA-I sequences. Using the recently described full-length consensus structure (Melchior et al. (2017)), we examined the structural properties of apoA-I that contribute to minimize the exposure of its constituent APRs. In silico saturation mutagenesis analysis of apoA-I demonstrated that an evolutionary-conserved APR, located between residues 14-19, contributes to the thermodynamic stability of the N-terminus and revealed a common destabilizing effect for amyloid-associated variants. Using molecular dynamics simulations, we studied the conformational and dynamical impact of five different amyloid variants on the structure of full-length apoA-I. Altogether, our results suggest that APR1 is a structural component that contributes to the stability of apoA-I helix bundle and emphasizes the destabilizing effect of amyloid variants, which is linked to subsequent APRs exposure in the case of G26R variant. This information is relevant to understand how a marginally stable, but metabolically active protein manages to initiate the formation of an amyloid structure and develop a severe pathology.

Results

Molecular evolution of apoA-I aggregating regions within the Sarcopterygii group

Given that apoA-I has four previously characterized aggregation-prone regions (APRs), we asked if this amyloid regions could be relevant to the protein functionality in spite of its know pathogenic role (Louros et al. 2015). To tackle this question, first of all we decided to explored the evolutionary conservation of these motifs within apoA-I sequences of sarcopterygian organisms. Our analysis was restricted to this group because of the large evolutionary distance between fishes and tetrapods, a factor that could mislead our results. We collected our sequences from the Ensembl database and constructed multiple sequence alignment (MSA) in order to identify the APRs present in other species based on the reported APRs for the human species. Then, we employed the Tango software to predict the sequence-based aggregation propensity of each one of the APRs sequences and also computed the sequence conservation from the MSA based on the Shannon entropy (H). Our amyloidogenicity results for the four APRs suggest that the APR1 (residues 14 to 19)

present an amyloid behaviour in more than 60% of the sequences of our dataset, followed by the APR4 in approximately 30% of the sequences. On the other hand, APR2 and APR3 presented a non aggregating behaviour in virtually all the sequences (Figure 1A). Regarding sequence conservation of aggregating regions, the data showed the sequence entropy of APRs residues are significantly higher than the average H value for apoA-I (P value = 1.84 x 10), implying that these amyloid regions are evolutionary less conserved (Figure 1B). The conservation across the different APRs seem to be different, with APR1 being more conserved (Supplementary figure 1).

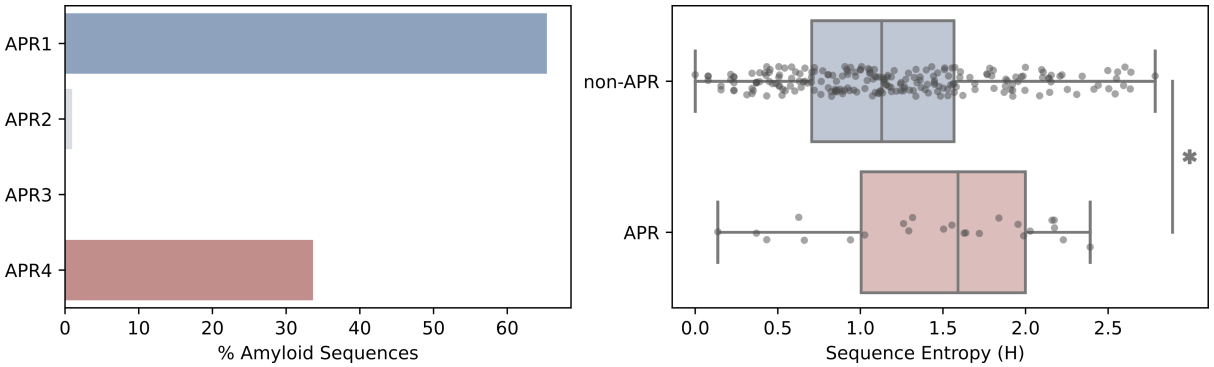

Figure 1: Figure 1 Evolutionary conservation of APRs within apoA-I sequences. **A** Percentage of sequences in our dataset that are described are amyloidogenic according to Tango (average score over 5%). **B** Sequence entropy (H) calculated for each residue inside the corresponding APR.

We decided to study the selection constraints affecting apoA-I as a way to provide further evidence of the conservation of its APRs. A maximum likelihood phylogeny reflecting the evolutionary relationships between sequences was reconstructed from the MSA (Supplementary figure 2A). Using this phylogeny as framework, we computed the site-wise evolutionary rates at codon level (dN/dS, ratio of nonsynonymous to synonymous mutations) and evaluated its statistical significance in order to evidence the presence of selection constraints acting on the apoA-I sequence. In particular, we employed different methods from the HyPHY package in order to detect both pervasive and episodic selection events. In general terms, the evolutionary rate profile of apoA-I revealed that most of the protein sequence displayed a dN/dS value significantly lower than 1 but this value tend to rise for the C-terminus of the protein (Supplementary figure 2B). Using a cartoon representation to depict the statistical evidence for the different types of selective pressure (negative, neutral or positive) at each site, we evidenced the extent of selection constraints, both pervasive and episodic, acting on apoA-I sequence. In accor-

dance with the entropy results, the residues corresponding to the APRs showed evidence of both purifying and neutral selection, meaning that some APR residues tend to be conserved during evolution but others can accept substitutions (Supplementary figure 2C). These data together supports the idea that some APRs, in particular APR1, have been conserved during the evolutionary history of apoA-I, while other APRs, like APR4, are much variable at the sequence level.

Comparative structural modelling of apoA-I historical sequences

Prompted by the sequence conservation of some of the APRs present in apoA-I, we decided to expand these results with information at the protein structure level. For this, we implemented a comparative modelling approach to compared several apoA-I structures corresponding to the ancestral sequences inferred from apoA-I phylogeny and several extant species, including reptiles, birds and mammals. To date, the most comprehensive and complete structure available for apoA-I is deposited on the webpage of the Davidson's lab ApoA-I consensus structure link, so we used it as the template for our homology-based modelling pipeline. We used Modeller to generate a structural model for each target sequence and the PyRosetta FastRelax tool to further refine it. A structure-based alignment of the best scoring model for each target sequence showed that the root mean square deviation (RMSD) between structures were in the range of 0.5-1, suggesting that the overall structure of apoA-I has been conserved along its evolutionary history (Supplementary figure 3A). Additionally, we computed the approximated intrinsic dynamic profile for each model using a gaussian normal network (GNM) model. These results showed that, besides the geometric properties of human apoA-I structure, its mean squared fluctuation (MSF) profiles are also conserved across ancestral and extant structures (Supplementary figure 3B).

We used these structural models to explore the intrinsic fluctuations levels and packaging number of the residue sites composing apoA-I APRs. Our results showed that APRs residue have significantly lower MSF values when compared with the value distribution for the non-APR residues of apoA-I (Figure 2A, P value = X). In a similar trend, the weighted contact number (WCN), a measure of how crowned is the molecular environment of a residue, also showed that APRs sites are consistently surrounded by a larger number of residues than the non-APR site of apoA-I (Figure 2B, P value = Y). Together, this

data suggest that apoA-I has conserved its structure and dynamics behaviour during its evolution. In this structural context, APRs residues are integrated into relatively rigid and densely packaged segments of apoA-I, a hallmark of functionally relevant sites for the protein structure (**cita?**).

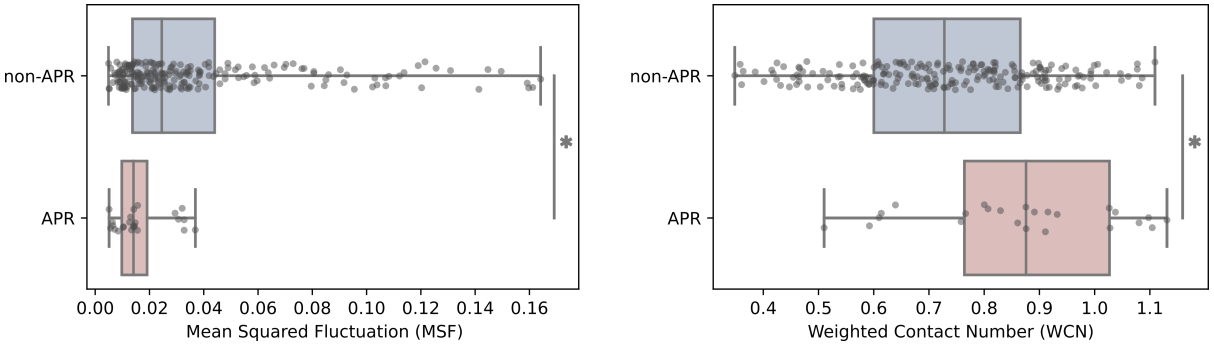

Figure 2: Figure 2 Dynamic and structural properties of APRs within apoA-I structures. **A** Fluctuation. **B** Contact number.

Amyloid-associated variants have a destabilizing effect on apoA-I monomer structure

In order to better understand the structural consequences of amyloid variants on apoA-I monomer, we explored their thermodynamic and pathological effects using in silico saturation mutagenesis. Destabilizing effect of each possible mutation in apoA-I sequence, represented by the difference in free energy ($\Delta\Delta$G) between wild type and mutant structure, was measured using the FoldX empirical force field and the MutateX automation pipeline. To complement this approach, variant impact on protein function was estimated using Rhapsody. We noticed from the $\Delta\Delta$Gs distribution that most of the variants had a moderate impact on apoA-I stability (-1 kcal/mol $<$ $\Delta\Delta$G $<$ 1 kcal/mol) (Figure 4A, complete FoldX results are available with Supplementary Figure 3). Further examination revealed that apoA-I structure is highly sensitive to mutations in the region of residues 7-28, which comprises the APR1 (Figure 4B). Rhapsody predictions also support this region as a mutation-sensible segment of apoA-I structure (Supplementary Figure 4). This result suggests that conservation of APR1 in apoA-I could be necessary to maintain the marginal thermodynamic stability of the -helix bundle despite the risk to undergo aggregation. In line with our observations, APRs have been recently proposed to play a stabilizing role on protein structure (Langenberg et al. 2020).

We used $\Delta\Delta$G values to highlight differences between pathogenic variants associated with

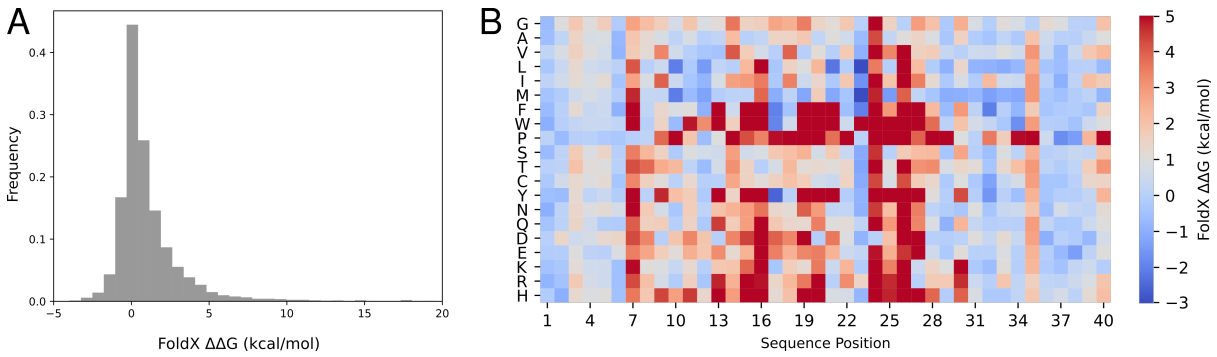

Figure 3: Figure 3 APR1 contributes to the stability of the -helix bundle in apoA-I. The protein structural stability was quantified using the FoldX engine. The free energy difference (ΔΔG) was calculated by comparison between the ΔG of the mutant and wild type sequence A ΔΔG values distribution corresponding to all possible mutations. B Heatmap of ΔΔG values for the first 40 residues of apoA-I N-terminal region.

amyloidosis or HDL deficiencies (Gogonea 2016), and natural variants reported by the gnomAD project (Karczewski et al. 2020). Our results evidenced that amyloid mutations had a destabilizing effect and a pathogenicity score significantly greater when compared with natural or HDL-deficiency variants (Figure 5A and 5B), emphasizing the relationship between structural destabilization and amyloid pathology onset. An interesting observation from this result is that HDL-deficiency mutations have similar effects compared with natural variants, suggesting that this type of mutations exert its pathogenic effect without disrupting apoA-I monomer stability. Given the fact that a small group of variants in the gnomAD database showed an elevated impact on protein stability (> 2 ΔΔG kcal/mol), we decided to investigate how frequently they occur at population level. Frequency spectrum (Figure 5C) showed that variants with a severe impact on protein stability were present at low frequencies, thus minimizing their deleterious effect on the population. In contrast, variants with the higher frequency in our dataset had a nearly neutral effect on stability. It is worth noting that although gnomAD excluded subjects with mendelian and pediatric diseases from its cohorts, we cannot rule out the possibility that some of these destabilizing variants correspond to non diagnosed pathologies.

Molecular dynamics simulations of apoA-I mutants

To complement our previous results showing the destabilizing effect of amyloid variants, we decided to study the dynamic properties of apoA-I amyloid mutants by conducting coarse-grain molecular dynamics simulations under the SIRAH force field. We selected

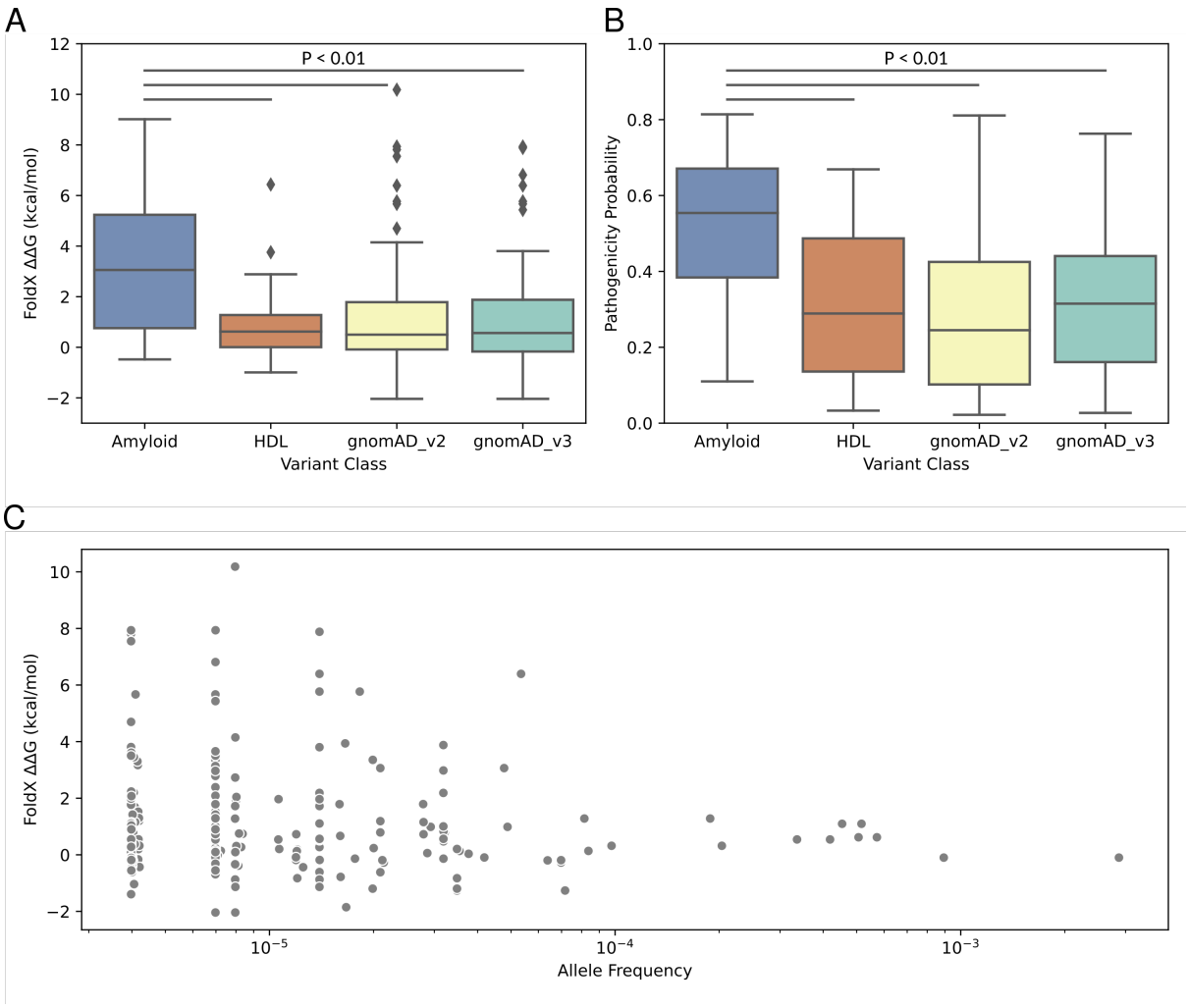

Figure 4: Figure 4 Impact of apoA-I variants on protein stability and function A-B Free energy difference ($\Delta\Delta G$) and Rhapsody pathogenicity distributions for amyloid, HDL and gnomAD variant classes (p-value < 0.01, Mann-Whitney rank test). C Allele frequency distribution for gnomAD variants against its predicted effect on protein stability.

four amyloid mutants (G26R, L60R, Δ107 and R173P) previously characterized by our group (Gaddi et al. 2020; Gisonno et al. 2020; Ramella et al. 2012; Rosú et al. 2015), plus the wild type protein, to prepare our simulation systems. Our selection also ensured that mutations were distributed throughout the apoA-I sequence.

In the first place, we explored the overall dynamics of our systems by means of its root mean square deviation (RMSD). The recently described consensus structure for apoA-I was used as reference coordinates for RMSD calculations. We observed a great variability in the RMSD values for all simulated systems (5.4-10 Å) during our 1 µs simulation, which could be related with the highly dynamic and marginally stable structure proposed for apoA-I. We did not evidence any significant differences between systems (Supplementary Table 2), suggesting that the impact of point mutations is negligible when compared against the intrinsic backbone dynamics.

Given the structural variability evidenced by RMSD, we decided to compute MD observables over the last 100 ns of the simulations. Position-specific root mean square fluctuations (RMSF) for each of the systems studied showed that loop regions 120-150 and 180-200 are the most flexible regions in apoA-I, while the N-terminal -helix bundle maintained a more compact structure during the simulation time (Supplementary Figure 5). These results are in good agreement with the MSF values computed by the GNM model (Figure 3C), reinforcing the dynamic profile obtained for apoA-I. The similar fluctuation profiles between the wild type apoA-I and the above-mentioned mutants suggest that mutations do not introduce major structural changes, at least during the simulation time frame. Because of this, we decided to compute more specific structural properties over the last 100 ns of the simulations.

To further characterize the structural impact of single point mutations on each system we measured the gyration radius (Rg) as a general descriptor of the protein shape for each system. When mutants were compared against the wild type system (Figure 6A), only the L60R system displayed a significantly higher Rg value, indicating a more extended conformation for this mutant. Mutants G26R and R173P also showed a tendency to present greater Rg values when compared against wild type, but they were statistically not significant, in part due to the highly variable nature of the apoA-I system.

We explored the possible role of mutations in amyloid aggregation of full length apoA-I by analyzing the solvent accessible surface area (SASA) of each APR in our five systems.

We noted a significant increase in the solvent exposure of the APR1 in the G26R system when compared against the wild type, while the other systems did not exhibit a significant increase of SASA values for any of the APRs (Figure 6B). Visualization of the final time frames of the trajectory corresponding to the G26R system showed a partial unfolding of the -helix bundle, which explains the increased exposure of APR1 (Figure 6C). Additionally, the G26R mutant evidenced the transitory formation of -strand secondary structures at the APR3. The low impact of the L60P, Δ107 and R173P variants on APRs exposure suggest that these mutants could require further post-translational modifications in order to undergo amyloid aggregation.

Figure 6 Molecular dynamics simulations of full-length apoA-I mutants A Gyration radius (Rg) of each system computed over the last 100 nanoseconds from five independent simulations. The L60R mutant showed a higher Rg when compared against the wild type system (p-value = 0.05, Student's Test). B Solvent accessible surface area (SASA) calculated for the APR1 (residues 14-19). The G26R system displayed a higher APR1 exposure when compared with the wild type system (p-value $<$ 0.05 Student's Test). C Graphical representation of the consensus model (WT) and the final snapshot of one of the replicas simulated for the G26R mutant. The substitution of glycine by arginine in position 26 destabilizes the helix bundle, expelling helix H6 with the concomitant solvent exposure of APR1 (left inset). A 180^{o} view rotation shows the -sheet hairpin formed between residues S224 and A232, corresponding to the APR3 (right inset).

Discussion

Molecular mechanism of amyloid aggregation for apoA-I remains largely unknown, due in part to the limited structural information given its inherent conformational plasticity (Gursky and Atkinson 1996). This work builds upon evolutionary, dynamical and structural features of apoA-I in order to provide a comprehensive characterization of the amyloid phenomena in this protein, complementing the extensive experimental results available. Collectively, our results suggest an intimate relationship between aggregating regions and structural stability in apoA-I. Additionally, MD simulations of full-length apoA-I mutants shed light on the first steps of the aggregation process in amyloid mutants.

First, we aimed to complement the few studies available that tackle the evolutionary history of apoA-I and its implications on its structure (Bashtovyy et al. 2010). An important observation that emerges from our results is that apoA-I evolution is tightly linked with the biophysical properties imposed by its constituent amphipathic -helices. This is especially evident in the case of prolines and positively charged residues, which are critical for apoA-I function and structure. Prolines have been extensively characterized as a fundamental component for apoA-I flexibility and stability, as their positioning at the beginning of the 22-mers induces a relative break of one helical segment respect to the other, allowing the protein rearrangement required for lipid removal and dynamic interactions with membranes and proteins interactors (Klon 2002 JMB https://doi.org/10.1016/S0022-2836(02)01143-9). In the case of charged residues, a strong lipid affinity has been attributed to the cationic residues within the polar face of the amphipathic -helices (Fuentes et al. 2018), as they interact with the negative heads of the phospholipids at the surface of lipid bilayers through a process designated "snorkeling" (Leman, Maryanoff, and Ghadiri 2013; Oda 2017). Also, arginine residues present at the helix 6 (R149, R153, and R160) are relevant for apoA-I-mediated activation of LCAT (Roosbeek et al. 2001 JLR PMID: 11160363). In a different trend, conservation of specific leucine residues could be driven by its involvement in lipid binding (Hovingh 2003 J Am Coll Cardiol 10.1016/j.jacc.2004.06.070, Fotakis 2013 JLR 10.1194/jlr.M038356) and stabilization of the hydrophobic clusters inside helix bundle (Gursky 2012 Biochemistry). In addition, given that fast evolving regions of proteins have been associated with greater flexibility (Tiwari 2018 https://doi.org/10.1016/j.sbi.2017.12.001; Campitelli et al. 2020), the higher evolutionary rate observed for the C-terminal region could be linked with the maintenance of the flexibility required for its lipid-binding properties.

The fact that apoA-I has conserved an aggregating segment (APR1) consistently along its evolutionary history raises questions about its structural relevance. Amyloid motifs have been proposed to contribute to protein structural stability through extensive interactions inside protein hydrophobic cores (Tartaglia et al. 2010 https://doi.org/10.1016/j.jmb.2010.08.013; Langenberg et al. 2020), which establish a trade-off between protein environment, foldability and aggregation propensity (Linding 2004 JMB https://doi.org/10.1016/j.jmb.2004.06.088; Monsellier 2008 PLoS CB https://doi.org/10.1371/journal.pcbi.1000199). Based on its conserved nature and

FoldX stability results, it is possible to hypothesize that APR1 is necessary to ensure the marginal stability of apoA-I -helix bundle, even though this region could trigger aggregation upon solvent exposure or proteolytic cleavage (Arciello FEBS Letters 2016 https://doi.org/10.1002/1873-3468.12468). Moreover, the presence of APR2 exclusively in human species represents a synergizing factor that could aggravate the amyloid behaviour of apoA-I, as demonstrated recently for the N-terminal peptide (Wong et al. 2012 10.1016/j.febslet.2012.05.007; Mizuguchi et al. 2019). In this context, the structural features of APR1 (low intrinsic flexibility and highly packaged environment) are likely to control its exposure to solvent and prevent aggregation events. Hydrogen-deuterium exchange experiments (Das et al. 2016) support the highly packaged nature of the -helix bundle and the low solvent exposure of APR1 in apoA-I.

Amyloidogenic variants are primarily located in the N-terminal region of apoA-I, whereas variants associated with HDL deficiencies are clustered in the H5-H7 region (Gogonea 2016; Matsunaga et al. 2010). Through a comprehensive evaluation of the destabilizing effect and pathogenicity of each possible mutation affecting apoA-I we demonstrated that amyloid variants have a significant destabilizing effect on the monomer structure. The fact that TANGO aggregation tendency of APRs was not modified by the introduction of amyloid mutations, supports the hypothesis that aggregation propensity per se has a limited impact on the aggregation process of full-length apoA-I (Raimondi et al. 2011). On the other hand, variants associated with HDL defects have a minimal effect on structural instability, which provides evidence that this type of disorders could be caused by mechanisms less dependent on protein unfolding and probably involving the disruption of interaction sites with protein interactors during the reverse cholesterol transport pathway. In addition, we believed that $\Delta\Delta G$ values derived from our in silico saturation mutagenesis would be useful as a proxy for the initial study of novel apoA-I mutants.

Taking advantage of the recently described consensus model of apoA-I (Melchior et al. 2017), our MD simulations of mutant G26R revealed a partial unfolding of the N-terminal -helix bundle and a significant increase in the exposure of APR1, which is also congruent with the destabilizing effect predicted from our $\Delta\Delta G$ calculations. This partial unfolding is in line with the experimental reports of increased susceptibility to proteases (Adachi et al. 2012) and greater hydrogen-deuterium exchange rate of the -helix bundle

(Das et al. 2016) for this mutant. Moreover, -sheet secondary structures present at APR3 could provide a template for the aggregation of full-length apoA-I (Das et al. 2014).

Altogether, our results obtained from full-lenght protein support the current hypothesis that unfolding of the helix bundle and exposure of aggregating regions represents the first steps of apoA-I-mediated amyloidosis (Mizuguchi et al. 2019). The mild effect of L60R, Δ107 and R173P variants on apoA-I structure and APRs exposure suggest that further modifications could be required to promote protein aggregation of these mutants, like oxidation or proteolytic cleavage (Witkowski 2018 10.1096/fj.201701127R; Chan 2015 10.1074/jbc.M114.630442).

Recently, the connection between the pro-inflammatory microenvironment and the formation of aggregation-prone species has been deeply characterized, reinforcing this hypothesis (Gisonno et al. 2020). Moreover, the presence of the N-terminal proteolytic fragment (residues 1-93) within patients' lesions raises the hypothesis that mutations may facilitate the cleavage of apoA-I by circulating proteases (Cavigiolio and Jayaraman 2014; Kareinen et al. 2018).

In agreement with the late onset of the hereditary apoA-I amyloidosis in patients, it may be hypothesized that mild chronic events may be required to induce the protein unfolding.

Materials and Methods

Evolutionary analysis of apoA-I sequences

A comprehensive dataset of sequences was generated by collating apoA-I orthologs available at Ensembl and Refseq databases (O'Leary 2015 Nucleic Acids Research https://doi.org/10.1093/nar/gkv1189; Yates 2020 Nucleic Acids Research https://doi.org/10.1093/nar/gkz966). To exclude low quality data, only sequences which did not contain ambiguous characters, had a proper methionine (M) starting codon and were longer than 200 amino acids were kept. Additionally, as both Ensembl and Refseq have overlapping data for some species, CD-HIT clustering tool (Fu et al. 2012) was employed to generate groups of similar sequences with an identity cut-off value of 0.98. Our fInal dataset comprised 215 protein sequences covering both Actinopterygii and Sarcopterygii lineages of Vertebrata. In order to reconstruct a maximum likelihood phy-

logeny, a multiple sequence alignment (MSA) was built from the protein sequences using ClustalO with default parameters (Sievers et al. 2011) and the phylogenetic inference was carried out with the IQ-TREE software (Minh et al. 2020). The substitution model was selected based on the ModelFinder evolutionary model fitting tool (Kalyaanamoorthy et al. 2017) and the ultrafast bootstrap implemented in IQ-TREE was used to calculate the support values for phylogeny branches (Minh, Nguyen, and Haeseler 2013). We rooted our phylogeny using cartilaginous fish species as outgroups, as proposed by de Carvalho et al. (2019). Visualization of the resulting phylogeny was carried out using the iTOL server (Letunic and Bork 2019).

Selective pressure acting on apoA-I sequence

Nucleotide coding sequences were retrieved for each protein in our dataset using the NCBI Entrez eutils tools for Refseq sequences and the Ensembl orthologs dataset. Because the evolutionary rate estimation requires a codon-level alignment, the software PAL2NAL was used to align codons in nucleotide sequence using a protein alignment as a guide (Suyama, Torrents, and Bork 2006). The Hypothesis Testing using Phylogenies (HyPhy) package was used to conduct evolutionary analysis on the codon-based alignment. Before testing for evidence of selective pressure, we conducted a recombination analysis using the Genetic Algorithm Recombination Detection (GARD) method, (Pond et al. 2006) in order to screen for possible recombination events in our alignment; it is known that the presence of recombination leads to a larger number of false positives in selection analysis. We inferred the natural selection strength (Omega, dN/dS) for each alignment position using our phylogeny as framework. We employed the Fixed Effects Likelihood (FEL) (Pond and Frost 2005) and the Fast Unconstrained Bayesian Approximation (FUBAR) methods (Murrell et al. 2013) to quantify the dN/dS ratio for each codon in the alignment. Although both methods provide similar information, FEL provides support for negative selection ($dN/dS < 1$) whereas FUBAR has more statistical power to detect positive selection ($dN/dS > 1$). Because codon alignment positions are difficult to put in structural context, data were extracted for codons occurring in wild type human apoA-I. Additionally, we estimated evolutionary rates based on alignments at amino acid levels using LEISR.

Coevolving residue pairs

Pairs of residues that are evolutionary correlated (coevolving sites) are useful to predict structural contacts. However, the repetitive structure of apoA-I and the lack of several ortholog sequences poses difficulties for this kind of analysis. To overcome these difficulties, putative coevolving residues were computed using the RaptorX server (Wang et al. 2017). RaptorX applies an ultra-deep convolutional residual neural network to predict contacts and distance and works particularly well on proteins without many sequence homologs. This method works by predicting the contact/distance matrix as a whole instead of predicting one residue pair independent of the others. RaptorX output represents the probability of two residues being in contact (i.e., their distance falling in the range 0-8 Å). Only residue pairs with a contact probability greater than 0.5 were retained.

Structural features measurement

Residue solubility profile for apoA-I consensus structure was computed with the CamSol method (Sormanni 2015 JMB http://dx.doi.org/10.1016/j.jmb.2014.09.026). CamSol first calculates an intrinsic solubility score for each residue, based only on sequence information. Then, the algorithm applies a score correction to the solubility profile from the previous step to account for the spatial proximity of amino acids in the three-dimensional structure and for their solvent exposure.

Fibril-forming segments were identified with the ZipperDB resource (https://services.mbi.ucla.edu/zippe

Fibrillation propensity is calculated as proposed by Thompson et al. (PNAS https://doi.org/10.1073/pnas.0511295103). Briefly, each hexapeptide not containing a proline from the query sequence is mapped onto the cross-beta crystal structure of the fibril-forming peptide NNQQNY. Energetic fit is evaluated with the RosettaDesign software (Kuhlman 2000 PNAS https://doi.org/10.1073/pnas.97.19.10383). Hexapeptides with energies below the threshold of -23 kcal/mol were considered as highly propense to fibrillation. Packaging level for residue i was represented by its Weighted Contact Number (WCN), which was calculated as follows:

Where, rij is the distance between the geometric center of the side-chain atoms for residue i and residue j. Calculations were carried out using a custom script developed by Sydykova et al. (2018 F1000 https://doi.org/10.12688/f1000research.12874.2).

Protein intrinsic dynamics was characterized using a coarse-grained simulation model based solely on protein topological information represented as a Gaussian Network Model (GNM). In this approach, protein structure is modelled as a network of nodes (alpha carbons) connected by springs. Numerical resolution of this model allows the calculation of the equilibrium displacement for all nodes (Mean Square Fluctuation, MSF), describing the global motions of the system. The ProDy package (Bakan, Meireles, and Bahar 2011) was used to adjust a GNM to the apoA-I consensus structure. We selected the first ten slow modes for analysis and plotting, since they have been reported previously as the main determinants of the global dynamics of protein structure (Kitao 1999 https://doi.org/10.1016/S0959-440X(99)80023-2).

Conservation of Aggregation Prone Regions (APRs)

Signal peptide sequences were trimmed and removed from the MSA to retain only the mature protein sequence. TANGO software (Fernandez-Escamilla et al. 2004) was used to detect APRs in the protein sequences dataset. This algorithm predicts beta-aggregation using a space phase where the unfolded protein can adopt one of five states: random coil, alpha-helix, beta-turn, alpha-helical aggregation or beta-sheet aggregation. Importantly, TANGO is based on the assumption that the core regions of an aggregate are fully buried. Predictions were carried out using default settings: no protection for the C-terminus and N-terminus, pH 7, temperature of 310° K and ionic strength of 0.1. Output files provide an aggregation score per position; as suggested in the TANGO manual and elsewhere, contiguous regions comprising five or more residues with a score of at least five were annotated as an APR. To address the impact of single point mutations in apoA-I aggregation tendency we ran TANGO for each mutant sequence and compared the scores profile against the wild type sequence. Sequence logos of each APR were plotted using the LogoMaker package (Tareen and Kinney 2019). TANGO software was downloaded from http://tango.crg.es using an academic license.

Thermodynamics impact of missense variants

The FoldX engine (Guerois, Nielsen, and Serrano 2002) implements an empirical energy function based on terms significant for protein structure stability. The free energy of unfolding (ΔG) of the protein includes terms for van der Waals interactions, solvation

of apolar and polar residues, intra and intermolecular hydrogen bonds, water bridges, electrostatic interactions and entropic cost for fixed backbone and side chains. Changes in free energy of folding upon mutation is calculated as the difference between the folding energy ($\Delta\Delta$G) estimated for the mutants and the wild type variants. Although FoldX seems to be more accurate for the prediction of destabilizing mutations and less accurate for the prediction of stabilizing mutations, in both cases it was shown that FoldX is a valuable tool to infer putative relevant sites for structural stability. FoldX 5 suite was downloaded from http://foldxsuite.crg.eu/academic-license-info.

We employed MutateX software (Tiberti et al. 2019) to automate the prediction of $\Delta\Delta$Gs associated with the systematic mutation of each available residue within apoA-I, by employing the FoldX energy function. At the heart of MutateX lies an automated pipeline engine that handles input preparation and performs parallel runs with FoldX. Basic steps involve protein data bank (PDB) structure repair (involving energy minimization to remove unfavorable interactions), model building for the mutant variants, energy calculations for both mutant and wild type structures and summarizing the estimated average free energy differences.

Pathogenicity scoring or missense variants

The Rhapsody prediction tool (Ponzoni et al. 2020) consists of a random forest classifier that combines sequence, structure, and dynamics-based features associated with a given amino acid variant and is trained over a comprehensive dataset of annotated human missense variants. Dynamical features include: mean-square fluctuations of the residue at the mutation site, which estimates local conformational flexibility; perturbation-response scanning effectiveness/sensitivity, accounting for potential allosteric responses involving the mutation site, and the mechanical stiffness at the sequence position of the mutated residue. These properties are computed from Elastic Network Models (ENM) representations of protein structures that describe inter-residue contact topology in a compact and computationally-efficient format that lends itself to a unique analytical solution for each structure. The algorithm was recently upgraded to include coevolutionary features calculated on conserved Pfam domains, and the training dataset was further expanded and refined. The latter combines annotated human variants from several publicly available datasets (Humvar, ExoVar, predictSNP, VariBench, SwissVar,

Uniprot's Humsavar and ClinVar). All analyses were performed using the Rhapsody server http://rhapsody.csb.pitt.edu/

Molecular Dynamics Simulations

Coarse grained Molecular Dynamics simulations were performed with the SIRAH force field (Machado et al. 2019) and GROMACS 2018.4 software package (Abraham et al. 2015). We employed the consensus model of human apoA-I in its monomeric and lipid-free state, proposed by Davidson et al. (Melchior et al. 2017). The PDB file was downloaded from Davidson Lab homepage (http://homepages.uc.edu/~davidswm/structures.html). Mapping atomic to coarse-grained representations was done with a Perl script included in SIRAH Tools (Machado and Pantano 2016). G26R, L60R, R173P and Δ107 mutants were generated with Chimera (Pettersen et al. 2004), editing the coordinates of the consensus model pdb file. For the case of the deletion mutant, we removed Lys107 and connected residues Lys106 and Trp108 with an unstructured segment using Modloop (Fiser and Sali 2003). Wild type apoA-I and the mutant systems were assembled using the following setup: The protein was placed inside an octahedron simulation box defined by setting a distance of 1.5 nm between the solute and the edges of the box. Systems were solvated setting a 150 mM NaCl concentration following the protocol proposed by Machado et. al. (2020). Energy minimization and heating steps were done following the protocol recommended by Machado et al. (2019) using positional restraints in the protein backbone to ensure side-chain relaxation, especially in the mutant models. Production runs were performed by quintuplicate in the absence of any positional restraint, generating 1 s trajectories at 310 K using a 1 bar NPT ensemble. Structural analysis was performed with GROMACS tools gmx rmsf, gmx gyrate and gmx sasa. Root mean square fluctuation was calculated for each residue aligning the full trajectory APOA-1 coordinates with the initial models. Radius of gyration and Solvent accessible surface areas (SASA) were obtained averaging the values corresponding to the last 0.1 s of simulation. The SASA calculations were measured over three amyloid prone regions, comprising residues 14-19 (APR1), 53-57 (APR2) and 227-232 (APR3).

Code Availability

All Python packages used were installed through the Conda environment manager into a single environment. A requirements file is available in the repository of this project in order to install dependencies used in our analysis. The workflow manager Snakemake was used in the evolutionary analysis in order to gain reproducibility and consistency of the results (Koster and Rahmann 2012). The Snakefile and Python scripts used in this work are available at https://github.com/tomasMasson/APOA1_evolution. Statistical Analyses and Visualizations Scipy Python library was used for data manipulation and all statistical analyses [Scipy]. Statistical significance was determined using Mann-Whitney U Test for variant's impact comparison and Student's Test for MD observables. MD graphs are reported as means ± standard deviation derived from five independent experiments. All visualizations were prepared with the Seaborn library [Seaborn].

References

Chroni, Angeliki, Tong Liu, Irina Gorshkova, Horng-Yuan Kan, Yoshinari Uehara, Arnold von Eckardstein, and Vassilis I. Zannis. 2003. "The Central Helices of ApoA-i Can Promote ATP-Binding Cassette Transporter A1 (Abca1)-Mediated Lipid Efflux." *Journal of Biological Chemistry* 278 (9): 6719–30. https://doi.org/10.1074/jbc.m205232200.

Das, Madhurima, and Olga Gursky. 2015. "Amyloid-Forming Properties of Human Apolipoproteins: Sequence Analyses and Structural Insights." In *Advances in Experimental Medicine and Biology*, 175–211. Springer International Publishing. https://doi.org/10.1007/978-3-319-17344-3_8.

Gaglione, Rosa, Giovanni Smaldone, Rocco Di Girolamo, Renata Piccoli, Emilia Pedone, and Angela Arciello. 2018. "Cell Milieu Significantly Affects the Fate of AApoAI Amyloidogenic Variants: Predestination or Serendipity?" *Biochimica Et Biophysica Acta (BBA) - General Subjects* 1862 (3): 377–84. https://doi.org/10.1016/j.bbagen.2017.11.018.

Gursky, O., and D. Atkinson. 1996. "Thermal Unfolding of Human High-Density Apolipoprotein a-1: Implications for a Lipid-Free Molten Globular State." *Proceedings of the National Academy of Sciences* 93 (7): 2991–95. https://doi.org/10.1073/pnas.93.7.2991.

Louros, Nikolaos N., Paraskevi L. Tsiolaki, Michael D. W. Griffin, Geoffrey J. Howlett, Stavros J. Hamodrakas, and Vassiliki A. Iconomidou. 2015. "Chameleon 'Aggregation-Prone' Segments of apoA-i: A Model of Amyloid Fibrils Formed in apoA-i Amyloidosis." *International Journal of Biological Macromolecules* 79 (August): 711–18. https://doi.org/10.1016/j.ijbiomac.2015.05.032.

Lund-Katz, Sissel, and Michael C. Phillips. 2010. "High Density Lipoprotein Structure–Function and Role in Reverse Cholesterol Transport." In *Cholesterol Binding and Cholesterol Transport Proteins:*, 183–227. Springer Netherlands. https://doi.org/10.1007/978-90-481-8622-8_7.

Manthei, Kelly A., Dhabaleswar Patra, Christopher J. Wilson, Maria V. Fawaz, Lolita Piersimoni, Jenny Capua Shenkar, Wenmin Yuan, et al. 2020. "Structural Analysis

of Lecithin:cholesterol Acyltransferase Bound to High Density Lipoprotein Particles." *Communications Biology* 3 (1). https://doi.org/10.1038/s42003-019-0749-z.

Melchior, John T, Ryan G Walker, Allison L Cooke, Jamie Morris, Mark Castleberry, Thomas B Thompson, Martin K Jones, et al. 2017. "A Consensus Model of Human Apolipoprotein a-i in Its Monomeric and Lipid-Free State." *Nature Structural & Molecular Biology* 24 (12): 1093–99. https://doi.org/10.1038/nsmb.3501.

Mucchiano, Gerd I., Bo Häggqvist, Knut Sletten, and Per Westermark. 2001. "Apolipoprotein a-1-Derived Amyloid in Atherosclerotic Plaques of the Human Aorta." *The Journal of Pathology* 193 (2): 270–75. https://doi.org/10.1002/1096-9896(2000)9999:9999%3C::aid-path753%3E3.0.co;2-s.

Navab, Mohamad, Srinivasa T. Reddy, Brian J. Van Lenten, G. M. Anantharamaiah, and Alan M. Fogelman. 2009. "The Role of Dysfunctional HDL in Atherosclerosis." *Journal of Lipid Research* 50: S145–49. https://doi.org/10.1194/jlr.r800036-jlr200.

Obici, Laura, Guido Franceschini, Laura Calabresi, Sofia Giorgetti, Monica Stoppini, Giampaolo Merlini, and Vittorio Bellotti. 2006. "Structure, Function and Amyloidogenic Propensity of Apolipoprotein a-i." *Amyloid* 13 (4): 191–205. https://doi.org/10.1080/13506120600960288.

Rader, Daniel J., Eric T. Alexander, Ginny L. Weibel, Jeffrey Billheimer, and George H. Rothblat. 2009. "The Role of Reverse Cholesterol Transport in Animals and Humans and Relationship to Atherosclerosis." *Journal of Lipid Research* 50: S189–94. https://doi.org/10.1194/jlr.r800088-jlr200.

Rosenson, Robert S., H. Bryan Brewer, Benjamin J. Ansell, Philip Barter, M. John Chapman, Jay W. Heinecke, Anatol Kontush, Alan R. Tall, and Nancy R. Webb. 2015. "Dysfunctional HDL and Atherosclerotic Cardiovascular Disease." *Nature Reviews Cardiology* 13 (1): 48–60. https://doi.org/10.1038/nrcardio.2015.124.

Sipe, Jean D., Merrill D. Benson, Joel N. Buxbaum, Shu-ichi Ikeda, Giampaolo Merlini, Maria J. M. Saraiva, and Per Westermark. 2016. "Amyloid Fibril Proteins and Amyloidosis: Chemical Identification and Clinical Classification International Society of Amyloidosis 2016 Nomenclature Guidelines." *Amyloid* 23 (4): 209–13. https://doi.org/10.1080/13506129.2016.1257986.